Univerzita Karlova

Filozofická fakulta

Ústav informačních studií a knihovnictví

Diplomová práce

Bc. Jan Dobiášovský

Přibližná shoda znakových řetězců a její aplikace na ztotožňování metadat vědeckých publikací

Approximate equality of character strings and its application to record linkage in metadata of scientific publications

Praha 2020

Vedoucí práce: Mgr. Jan Dvořák, Dr.

Rád bych moc poděkoval Dr. Dvořákovi za cené rady, trpělivost, ochotu a skvělý odborný dohled během psaní této práce. Mé poděkování dále patří také mé rodině, přátelům a kolegům za jejich podporu.

Prohlášení:

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny využité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne Jméno a příjmení

……..…… ……..……

Klíčová slova (česky):

ztotožňování záznamů, přibližná shoda znakových řetězců, deduplikace, fuzzy matching, zpracování a klasifikace dat, správa databází, metadata publikací

Keywords (English):

record matching, approximate string matching, deduplication, fuzzy matching, data processing and classification, database management, publication metadata

Abstrakt:

Práce zkoumá použití přibližné shody znakových řetězců v procesu ztotožňování metadat vědeckých publikací. V teoretické části je poskytnut úvod do problematiky, popsán proces ztotožňování záznamů a pět používaných metrik pro vyjádření podobnosti znakových řetězců (Levenshteinova vzdálenost, Jarova vzdálenost, Jaro-Winklerova vzdálenost, kosinová vzdálenost q-gramů a Jaccardův koeficient). V praktické části je zkoumána možnost použití těchto metrik v systému V3S pro evidenci vědeckých publikací na ČVUT a jejich porovnání. Na trénovací množině byla potvrzena vhodnost využití v systému V3S a stanoveny optimální prahy pro jednotlivé metriky ma základě měr F_1, F_2 a F_3.

Abstract:

The thesis explores the application of approximate string matching in scientific publication record linkage process. An introduction to record matching along with five commonly used metrics for string distance (Levenshtein, Jaro, Jaro-Winkler, Cosine distances and Jaccard coefficient) are provided. These metrics are applied on publication metadata from V3S current research information system of the Czech Technical University in Prague. Based on the findings, optimal thresholds in the F_1, F_2 and F_3-measures are determined for each metric.

1. Úvod a cíle práce

Masivní objem dat produkovaný společností klade extrémně vysoké nároky na jejich efektivní správu a manipulaci s nimi. O jednom objektu reálného světa existuje mnoho dat. V případě, kdy jsou údaje o objektu pořizovány či zpracovávány více nezávislými způsoby, budou se úrovně kompletnosti těchto dat lišit. Dále se v datech mohou vyskytnout také nepřesnosti i chyby, a tím pádem klesá jejich informační hodnota. Nekvalitní data znesnadňují interpretaci vlastní informace a její zpracování. Zvýšení kvality dat je obtížná úloha, která vyžaduje relativně velké množství kvalifikované lidské práce. Zejména u středních objemů dat je čištění lidmi opravdu pomalé a drahé. V případě velkých objemů dat pak v podstatě nerealizovatelné. Aby bylo možné vzniklé chyby napravit nebo odstranit duplicitní záznamy, je nutné tento objekt ztotožnit s reálnou entitou, kterou reprezentuje. V případě, kdy neexistuje jednoznačný identifikátor pro daný objekt, je nutné použít alternativní metody pro propojení dat s objekty reálného světa, které by měly reprezentovat. Význačným typem takových objektů jsou vědecké publikace. Jejich metadata jsou evidována mnoha aktéry v průběhu výzkumu, v procesu publikování i po jeho dokončení.

Práce se zabývá testováním použití metod přibližné shody znakových řetězců při automatizovaném zpracování metadatových záznamů vědeckých publikací. Metody jsou aplikovány na institucionální systém V3S vyvinutý a používaný Českým vysokým učením technickým v Praze, který slouží pro evidenci aktuálních vědeckých výsledků, jejich odesílání do národní databáze RIV a pro jejich využití uvnitř organizace. Vedle ručního zadávání jsou metadata systému V3S získávána z různých externích zdrojů, z nichž hlavní jsou databáze Web of Science (WoS) a Scopus. Z těchto zdrojů jsou též získávány údaje o citujících publikacích. Situace, kdy jedna publikace je reprezentována v obou těchto databázích, je typická. Během integrace těchto záznamů v systému V3S je třeba tyto záznamy ztotožňovat.
Předpokládá se, že nasazení metod přibližné shody znakových řetězců pro systém V3S umožní tyto duplicity snáze detekovat, čímž usnadní slučování a zvyšování kvality informačního obsahu systému v případech, kdy není dostupný identifikátor DOI.

2. Ztotožňování záznamů

Obecně ztotožňování záznamů spočívá v porovnávání jejich polí. K tomu se v některých případech používá tzv. *přibližná shoda*. Ta umožňuje detekovat i znakové řetězce, které nejsou zcela shodné, ale jsou si navzájem *podobné*. Použití v praxi primárně spočívá v možnosti studovat vztahy mezi dvěma a více datovými elementy ze separátních souborů pomocí tvorby rámců, odstraňování duplikátů a případně kombinování souborů (Winkler, 1995).

Anglický název pro ztotožňování záznamů *record linkage* pochází ze stejnojmenného článku publikovaného Halbertem L. Dunnem již v roce 1946 (Dunn, 1946). Probabilistické základy moderního ztotožňování záznamů byly položeny o 13 let později Howardem Bordenem Newcombem (Newcombe, 1959) a formalizovány Ivanem P. Fellegim a Alanem B. Sunterem (Fellegi, 1969).

2.1 Fellegi-Sunterův model

Fellegi-Sunterův model popisuje počítačově orientované řešení rozpoznávání záznamů ve dvou souborech reprezentujících identickou osobu, objekt nebo událost. Tento stav se nazývá *shoda*. V rámci procesu probíhá porovnání dostupných charakteristik a hodnot ve dvou záznamech (každý z jednoho souboru) a jeho výsledkem je, zda došlo či nedošlo ke shodě, případně jaká je pravděpodobnost, že existuje shoda (tedy platí, že data v obou souborech reprezentují stejný objekt reálného světa).

2.1.1 Matematický model

Předpokládejme dvě množiny **A** a **B**, jejichž elementy jsou příslušně označeny ***a*** a ***b***.

Vytvoříme-li množinu uspořádaných dvojic z prvků z obou množin, tzv. karteziánský součin:

$$A \times B = \{(a, b), a \in A, b \in B\}$$

můžeme předpokládat, že tato množina je sjednocením následujících dvou množin:

$$M = \{(a, b); a = b, a \in A, b \in B\}$$

$$U = \{(a, b); a \neq b, a \in A, b \in B\}$$

Tedy některé elementy se mohou vyskytovat v obou množinách. Tyto množiny na základě tohoto vztahu nazýváme jako *propojené* a *nepropojené*.

Každá jednotka v množině záznamů má určitý počet charakteristik, které se na ni pojí (např. jméno, věk, pohlaví, rodinný stav, adresa, místo, datum narození). Předpokládáme existenci dvou zaznamenávacích procesů. Výsledkem těchto procesů je záznam pro každou jednotku obsahující vybrané charakteristiky (např. stáří nebo adresa v určitý den) a zároveň jsou generovány určité chyby a nedostatky (překlepy, nevyplněné položky). Jako výsledek vzniklý na základě těchto chyb mohou dvě nepropojené jednotky množin A a B vytvořit **identické** záznamy, nebo naopak dvě propojené jednotky mohou vytvořit **odlišné** záznamy. Takovéto záznamy korespondující s členy množin A a B označujeme ***α(a)*** a ***β(b)***.

Dále také předpokládáme, že jednoduché náhodné vzorky z množin **A** a **B** jsou vybrány z opačných množin (A_s = B a B_s = A), ale nevylučujeme možnost, že $\mathbf{A_s = A}$ **a** $\mathbf{B_s = B}$. Dané vstupní soubory L_A a L_B jsou považovány za výsledek aplikace procesu generujícího množiny A_a a B_b (Felegi, Sunter, 1969).

Prvním krokem v pokusu o ztotožňování záznamů v daných souborech (tedy identifikace záznamů, které korespondují s propojenými členy A a B) je porovnání obou záznamů. Výsledek je množina výrazů (kódů) reprezentujících shody jako např.: *"jména jsou stejná"*, *"jména jsou stejná, jméno je Karel"*; stejně tak neshody - *"jména nejsou stejná",* nebo možné shody - *"jméno chybí v jednom ze záznamů", "v rámci adresy jsou stejná města, ale odlišné adresy"*.

Formální zápis *porovnávacího vektoru* jako vektorové funkce záznamů ***α(a)*** a ***β(b)*** je:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], ..., \gamma^K[\alpha(a)]\}$$

γ (případně je možné zapisovat jako γ(a,b) nebo γ(*α,β)* **reprezentuje karteziánský součin množin A × B zmíněných výše**. Množina možných realizací γ se nazývá *porovnávací prostor* a je označována jako **Γ**.

V průběhu propojovací operace sledujeme γ (a,b) a chceme provést rozhodnutí, že:

a) *(a,b) je propojený pár (a,b)* $\in M$

   Toto rozhodnutí se nazývá **pozitivní propojení** (A_1)

b) *(a,b) je nepropojený pár (a,b)* $\in U$

Toto rozhodnutí se nazývá **pozitivní nepropojení** (A_3)

V některých případech se ale může stát, že nejsme schopni výše uvedená rozhodnutí na dané *úrovni důvěryhodnosti* provést (např. program může klasifikovat oba řetězce jako shodné, pokud panuje 90% shoda. Pokud je shoda nižší, nelze klasifikovat). V tomto případě můžeme povolit třetí rozhodnutí $\mathbf{A_2}$ nazývané **možné propojení.** První dvě možná rozhodnutí nazýváme **pozitivní dispozice**.

Propojovací pravidlo L lze nyní definovat jako mapování **Γ** (porovnávací prostor) na množinu rozhodovacích funkcí D = d(γ), kde platí:

$$d(\gamma) = \{P(A_1 \mid \gamma), P(A_2 \mid \gamma), P(A_3 \mid \gamma)\}; \gamma \in \Gamma$$

$$\sum_{i=1}^{3} P(A_i \mid \gamma) = 1$$

Ke každé pozorované hodnotě γ propojovací pravidlo přiřadí pravděpodobnosti pro každé ze tří možných rozhodnutí. Pro některé, nebo i všechny hodnoty γ může rozhodovací funkce degenerovat náhodnou proměnnou a tedy je jednomu z rozhodnutí přiřazena pravděpodobnost s hodnotou 1. Zároveň musíme zvážit míru chybovosti asociovanou s propojovacím pravidlem. Předpokládáme, že pár záznamů [α(a), β(b)] je vybrán pro porovnání na základě pravděpodobnostního procesu $L_A \times L_{\mathbf{B}}$, což je ekvivalentní k náhodnému výběru páru elementů (a,b) z A×B během konstrukce z L_A a L_B. Výsledný porovnávací vektor γ[a(α),b(β)] je generovaná proměnná. Podmíněnou pravděpodobnost výskytu konkrétních hodnot porovnávacího vektoru γ za předpokladu, že (a,b) ∈ M, označujeme jako **m(γ)**.

Tedy:

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in M\}$$
$$= \sum_{(a,b) \in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot [(a,b) \mid M]$$

Podobně můžeme podmíněnou pravděpodobnost výskytu konkrétních hodnot porovnávacího vektoru γ za předpokladu, že (a,b) ∈ U, označit jako **u(γ)**.

Tedy:

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)] \mid (a,b) \in U\}$$
$$= \sum_{(a,b)\in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot [(a,b) \mid U]$$

Existují dva typy chyb asociovaných s propojovacím pravidlem s následujícími pravděpodobnostmi:

a) Dva objekty, které nejsou stejná entita, jsou propojeny

$$P(A_1 \mid U) = \sum_{\gamma\in\Gamma} u(\gamma) \cdot P(A_1 \mid \gamma)$$

b) Dva objekty reprezentující stejnou entitu, nejsou propojeny

$$P(A_3 \mid M) = \sum_{\gamma\in\Gamma} m(\gamma) \cdot P(A_3 \mid \gamma)$$

Propojovací pravidlo **L** pro prostor Γ je považováno za propojovací pravidlo v úrovních **μ, λ** pro ***0 < μ < 1*** a **0 < λ <1**, když platí:

$$P(A_1 \mid U) = \mu$$
$$P(A_3 \mid M) = \lambda$$

Takové pravidlo označujeme jako **L(*μ*, λ, Γ)**.

V množině propojovacích pravidel v prostoru Γ, která splňují předchozí dvě podmínky, bude propojovací pravidlo L(*μ*, λ, Γ) považováno za *optimální propojovací pravidlo*, jestliže platí:

$$PP(A_2 \mid L) \leq P(A_2 \mid L')$$

pro všechna L'(*μ*, λ, Γ).

Optimální propojovací pravidlo tedy maximalizuje pravděpodobnosti pozitivních dispozicí v porovnání (tedy A_1 a A_3), která jsou podmíněna fixními pravděpodobnostmi chybovosti. Jinak řečeno, je minimalizována pravděpodobnost, že nebude možné vytvořit pozitivní dispozici a tedy nebude nutné rozhodnutí A_2, které vyžaduje náročné manuální operace, využívat více, než je nutné. Pokud pravidlo má pravděpodobnost jevu A_2 příliš velkou, je vhodné zvážit, zda je reálně použitelné. (Felegi, 1969)

2.1.2 Základní věta pro propojovací pravidla

Předpokládejme propojovací pravidlo L_0 na Γ. Začneme definicí unikátní řady konečné množiny možných realizací *γ*.

Pokud existuje taková hodnota γ, že m(γ) a u(γ) jsou rovny nule, poté je (bezpodmínečná) pravděpodobnost realizace γ rovna také nule. A v tom případě není nutné ji zahrnout do Γ. Nyní seřadíme náhodně všechny γ, pro která platí m(γ) > 0, ale u(γ) = 0, a zbylá γ tak, aby odpovídající posloupnost m(γ)/u(γ) byla nerostoucí. Pokud hodnota m(γ)/u(γ) je stejná pro více γ, lze tyto hodnoty řadit náhodně. Seřazenou množinu {γ} indexujeme indexem *i*; (i=1,2, … , N_Γ); a zapisujeme $u_i = u(\gamma_i)$; $m_i = m(\gamma_i)$.

Nechť (μ, λ) je pár přípustných úrovní chybovosti. Vybereme hranice sumace *n* a *n'* tak, aby platilo:

$$\sum_{i=1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i$$

$$\sum_{i=n'}^{N_\Gamma} m_i \geq \lambda > \sum_{i=n'+1}^{N_\Gamma} m_i$$

kde N_Γ je počet členů v Γ.

Předpokládejme, že došlo k naplnění výše uvedených předpokladů a platí $1 < n < n'-1 < N_\Gamma$. V tomto stavu je zajištěno, že úrovně chybovosti (μ, λ) jsou přípustné. Pro propojovací pravidlo $L_0(\mu, \lambda, \Gamma)$ je porovnávacímu vektoru γ_i přiřazeno:

1) rozhodnutí A_1 (pozitivní propojení), pokud $i \leq n-1$,

2) rozhodnutí A_2, pokud $n < i \leq n'-1$,

3) rozhodnutí A_3 (pozitivní nepropojení), pokud $i \geq n' + 1$.

Pro $i = n$ nebo $i = n'$ je nutné provést náhodné rozhodnutí, aby bylo možné přesně dosáhnout úrovní chybovosti μ a λ. Formálně zapsáno:

$$d(\gamma_i) = \begin{cases} (1,0,0) & i \geq n-1 \\ (P_\mu, 1-P_\mu, 0) & i = n \\ (0,1,0) & n < i \leq n'-1 \\ (0,1-P_\lambda, P\lambda) & i = n' \\ (0,0,1) & i \geq n'+1 \end{cases}$$

kde P_μ a P_λ jsou definovány pomocí následujících rovnic:

$$u_n \cdot P_\mu = \mu - \sum_{n=1}^{i=1} u_i$$

$$m_{n'} \cdot P_\lambda = \lambda - \sum_{i=n'+1}^{N_\Gamma} m_i$$

Věta (Fellegi, 1969):

Nechť $L_0(\mu, \lambda, \Gamma)$ je propojovací pravidlo definované úrovností chybovosti $d(\gamma_i)$. Pak L_0 je optimální propojovací pravidlo v porovnávacím prostoru Γ na úrovních (μ, λ).

Fellegi a Sunter také uvádí dva důsledky:

Důsledek 1 (Felegi, 1969):

Pokud:

$$\mu = \sum_{i=1}^{n} u_i, \quad \lambda = \sum_{i=n}^{N_\Gamma} m_i, \quad n < n',$$

je optimální propojovací pravidlo $L_0(\mu, \lambda, \Gamma)$ na úrovních (μ, λ) transformováno na:

$$d(\gamma_i) = \begin{cases} (1,0,0) & \text{pokud } 1 \leq i \leq n \\ (0,1,0) & \text{pokud } n \leq i \leq n' \\ (0,1,0) & \text{pokud } n' \leq i \leq N_\Gamma \end{cases}$$

Pokud definujeme:

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$

$$T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$$

můžeme propojovací pravidlo $d(\gamma_i)$ také zapsat:

$$d(\gamma) = \begin{cases} (1,0,0) \text{ pokud } T_\mu \leq m(\gamma)/u(\gamma) \\ (0,1,0) \text{ pokud } T_\lambda < m(\gamma)/u(\gamma) < T_\mu \\ (0,0,1) \text{ pokud } m(\gamma)/u(\gamma) \leq T_\lambda \end{cases}$$

Důsledek 2 (Felegi, 1969):

Nechť T_μ a T_λ jsou dvě kladná čísla. Pokud platí $T_\mu > T_\lambda$, pak existuje přípustná hodnota úrovní chybovosti (μ, λ) korespondujících s T_μ a T_λ , ve kterých je propojovací pravidlo $d(\gamma)$ nejlepší možné. Na těchto úrovních definovaných:

$$\mu = \sum_{\gamma \in \Gamma\mu} u(\gamma)$$

$$\lambda = \sum_{\gamma \in \Gamma\lambda} m(\gamma)$$

kde:

$$\Gamma_\mu = \{\gamma : T_\mu \leq m(\gamma)/u(\gamma)\}$$

$$\Gamma_\lambda = \{\gamma : m(\gamma)/u(\gamma) \leq T_\lambda)\}$$

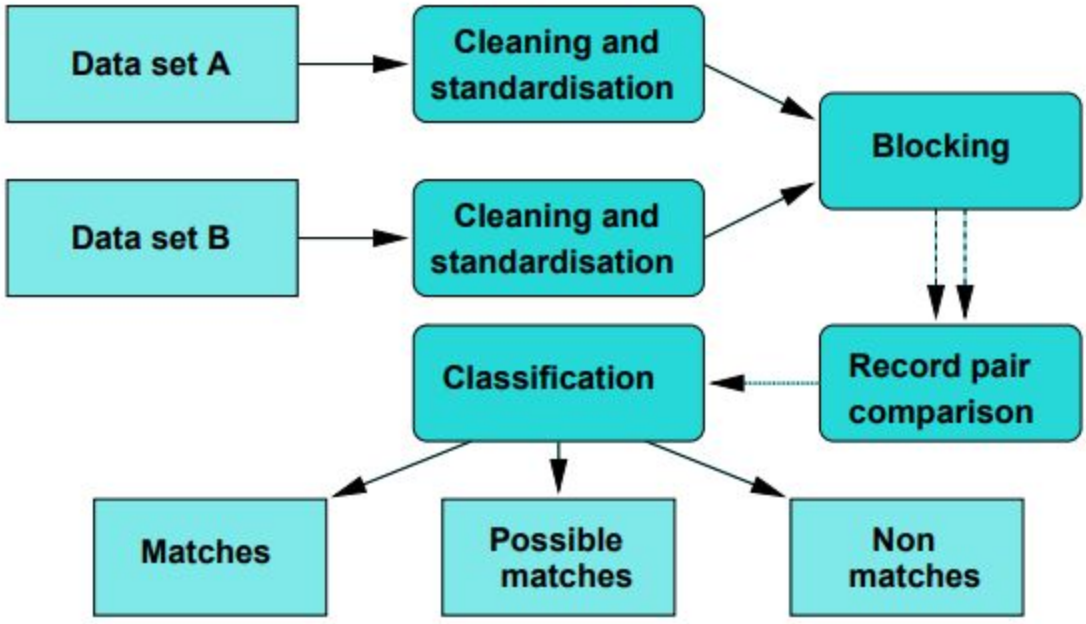
V mnoha praktických aplikacích můžeme tolerovat úrovně chybovosti dostatečně vysoké na to, aby se předešlo rozhodnutí A_2. V tomto případě používáme *n* a *n'* nebo T_μ a T_λ tak, že prostřední případ v d(γ) je prázdný. Jinými slovy, všechna (a,b) jsou přiřazena souborům M nebo U.

2.2 Proces ztotožňování záznamů

V této kapitole bude popsáno základní shrnutí procesu ztotožňování na základě publikace *"Data Matching, Data-centric Systems and Applications"* Petera Christena.

Christen (2012) ve své publikaci *"Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection"* rozděluje celý proces na následující části:

1) Čištění dat
2) Blocking
3) Porovnávání párů záznamů
4) Klasifikace párů záznamů
5) Evaluace propojovací kvality a komplexity



Obr. č. 1: Schéma procesu ztotožňování záznamů. Převzato z Christen (2007)

V následujících podkapitolách bude na základě této publikace a dalších dostupných zdrojů proces detailněji popsán.

2.2.1 Čištění dat

Před samotným procesem ztotožňování záznamů je nutné data určená k importu patřičně připravit, aby se minimalizovalo riziko chyb. Čištění dat (anglicky *data cleaning*) se zabývá detekcí, odstraňováním chyb a inkonzistencí z dat za účelem zlepšení jejich kvality. Příkladem problémů vyskytujících se v rámci jedné datové sbírky, jako je např. soubor nebo databáze, jsou překlepy a chybějící či nevalidní informace.

Problematikou čištění dat se zabývají Rahm a Do (2000). V článku *"Data Cleaning: Problems and Current Approaches"* klasifikují problémy s kvalitou dat, které jsou čištěním dat řešeny, a poskytují přehled přístupů k řešení těchto problémů.

Vzhledem k tomu, že čištění dat je poměrně náročný proces, prevence tzv. "špinavých" dat by měla být prioritou. K tomu je zapotřebí dobrý design databázového schématu a limitujících parametrů pro jednotlivé hodnoty atributů.

Tyto problémy lze rozlišit podle počtu zdrojů na jedno, či více-zdrojové, a podle úrovně, na které se vyskytují, na úroveň schématu databáze nebo výskytů (Rahm, Do, 2000). Stejný zdroj také popisuje příklady problémů s daty:

1) **Jednozdrojové problémy**
    a) **Úroveň schématu**
        i) Jedinečnost
        ii) Referenční integrita
        iii) …
    b) **Úroveň výskytu**
        i) Překlepy
        ii) Redundance / duplicity
        iii) Protichůdné hodnoty
        iv) ....
2) **Vícezdrojové problémy**
    a) **Úroveň schématu**
        i) Jmenné konflikty
        ii) Strukturální konflikty
        iii) …
    b) **Úroveň výskytu**
        i) Nekonzistentní agregace
        ii) Nekonzistentní časové formáty
        iii) ....

2.2.1.1 Jednozdrojové problémy

Kvalita dat zdroje hluboce závisí na úrovni pečlivosti kontroly schématem nebo integritními omezeními kontrolujícími povolené hodnoty. Pro zdroje bez schématu, jako jsou soubory, existuje minimum omezení, jaká data mohou být vložena a uložena, a tedy je zde vysoký risk chyb a inkonzistencí. Databázová řešení používají restrikce specifického datového modelu (např.

relační přístup vyžaduje jednoduché hodnoty atributů a referenční integritu) a další omezení specifické pro jednotlivé aplikace. Problémy na úrovni schématu se nejčastěji vyskytují buď kvůli limitům datového modelu, chybám v návrhu schématu, nebo nedostatečné kontrole omezujících podmínek pro jednotlivá pole. (Rahm, Do, 2000)

Pro obě úrovně můžeme rozlišit různé rámce problémů na atributy (pole), záznamy, typy záznamů a zdroje. Rahm (2000) uvádí i několik případů ilustrovaných v následujících obrázcích:

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = current year - birth year should hold
Record type	Uniqueness violation	emp_1=(name="John Smith", SSN="123456"); emp_2=(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

Table 1: Examples for single-source problems at schema level (violated integrity constraints)

Obr. č. 2: Příklady jednozdrojových problémů na úrovni schématu. Převzato z Rahm (2000)

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	$name_1$= "J. Smith", $name_2$="Miller P."	usually in a free-form field
	Duplicated records	emp_1=(name="John Smith",...); emp_2=(name="J. Smith",...)	same employee represented twice due to some data entry errors
	Contradicting records	emp_1=(name="John Smith", bdate=12.02.70); emp_2=(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

Table 2: Examples for single-source problems at instance level

Obr. č. 3: Příklady jednozdrojových problémů na úrovni výskytu. Převzato z Rahm (2000)

2.2.1.2 Vícezdrojové problémy

Problémy vyskytující se v jednotlivých souborech jsou násobeny při integraci více zdrojů do jednoho. Každý zdroj může obsahovat špinavá data, případně data v jednotlivých zdrojích mohou být duplikáty nebo si navzájem odporovat. Na úrovni schématu jsou rozdíly mezi datovými modely řešeny za pomoci překladu nebo integrací (zejména proto, že jednotlivé zdroje jsou nejčastěji vyvíjeny a poté udržovány za odlišných podmínek a za jiným účelem). Hlavní problémy v rámci designu schématu jsou pojmenování a strukturální konflikty. Konflikty v pojmenování vznikají, když je stejné jméno použito pro různé objekty (homonyma) nebo různá jména použita pro stejný objekt (synonyma). Strukturální konflikty se vyskytují v mnoha variacích a vztahují se na různé reprezentace stejných objektů v různých zdrojích (např. atribut vs. tabulka). Zároveň se na úrovni instance se můžeme setkat s různými reprezentacemi hodnot, i přestože datové typy a názvy atributů jsou stejné, jako např. značení měny nebo pohlaví). Informace mohou také být odlišné úrovní granularity. Je rovněž potřeba předpokládat, že v rámci integrace více zdrojů se může vyskytnout i řada jednozdrojových problémů (duplikáty, odporující údaje).

2.2.1.3 Proces čištění dat

Rahm a Do (2000) shrnují proces čištění do následujících fází:

1) ***Analýza dat***

   "Chceme-li detekovat, jaké typy chyb a nekonzistencí je potřeba odstranit, je zapotřebí provést detailní analýzu. Kromě provedení manuální inspekce dat nebo jejich části je vhodné použít programy pro analýzu za účelem získání metadat o vlastnostech dat a detekce problémů s kvalitou dat."

2) ***Definice transformačního workflow a mapovací pravidla***

   "V závislosti na počtu datových zdrojů, úrovni heterogenity a špinavosti jejich dat, může proběhnout velké množství kroků transformace a čištění dat. V některých případech je použit překlad schématu pro mapování zdrojů pro běžné datové modely, pro datové sklady se typicky používá relační reprezentace. První kroky v čištění mohou opravit jednozdrojové instanční problémy a připravit data pro integraci. Pozdější kroky řeší integraci na úrovni schématu a datového obsahu a vícezdrojové instanční problémy, jako

například duplikáty. Řešení pro problémy na úrovni schématu v rámci datových transformací a čistících postupů by měly být specifikovány deklarativními dotazy a mapovacím jazykem co nejpodrobněji, aby bylo umožněno automatické generování transformačního kódu. Dále by také mělo být možné spustit uživatelem psaný čistící kód a specializované nástroje během transformačního workflow. Kroky v rámci této transformace mohou žádat o uživatelskou zpětnou vazbu pro datové instance, kde chybí definovaná čistící logika."

3) ***Verifikace***

"Správnost a efektivita transformačního workflow a transformačních definic by měla být testována a hodnocena např. na části dat nebo jejich kopii za účelem zlepšení definic v případě potřeby. Může také dojít k vícenásobnému opakování analýzy, designu a ověřovacích kroků, vzhledem k tomu, že některé chyby se mohou projevit až po aplikaci transformace."

4) ***Transformace***

"Proces transformačních kroků buď spuštěním patřičného ETL workflow (extract, transform, load - proces kopírování) pro načítání a aktualizaci datového skladu, nebo transformace v rámci odpovědi na dotaz na více zdrojů."

5) ***Backflow čistých dat***

"Po eliminaci (jednozdrojových) chyb, by měla čistá data nahradit špinavá data v originálních zdrojích. Jednak za účelem poskytnutí vylepšených dat původní aplikaci, ale také zabránění nutnosti čistící proces provést znovu při opakovaných extrakcích dat."

2.2.2 Indexace

Jakmile databázové tabulky obsahují čistá data, je možné začít porovnávat záznamy. Je zřejmé, že velké procento veškerých porovnávacích operací bude prováděno nad záznamy, které nejsou propojené. Za účelem ušetření výpočetního výkonu a snížení časové náročnosti jsou aplikovány *indexační techniky*. Tyto techniky filtrují páry kandidátních záznamů, které mají menší šanci na propojení, a naopak generují kandidátní páry vhodnější k podrobnějšímu porovnání (existuje větší šance k propojení) (Christen, 2011).

Christen (Christen, 2011) a Baxter (Baxter, 2003) popisují několik různých metod indexace. Pro účely této práce budou popsány tři základní typy: blokace, indexace seřazených sousedů a indexace pomocí q-gramů.

2.2.2.1 Blokace

Záznamy jsou rozčleněny do bloků, které sdílí identický *blokovací klíč* (block key). Tento klíč je definován na základě společných atributů záznamů z jednotlivých datasetů jako např. stejný rok vydání, stejná čtyři začáteční písmena příjmení, stejné poštovní směrovací číslo apod. (Baxter, 2003). Blokace záznamů vyžaduje kompromis mezi náklady a přínosy (Christen, 2012). Pokud je blok příliš velký, může stále vyžadovat přílišný počet porovnávacích operací. Pokud je naopak příliš malý, může dojít k vyřazení záznamů, které lze propojit, a tím se sníží přesnost propojovacího procesu. Například použití pohlaví jako bloku rozdělí databázi pouze napůl, čímž se samozřejmě počet nutných porovnávacích operací zmenší na polovinu, ale u velkých datasetů to nemusí stačit. Naopak blokování pomocí rodného čísla databázi potenciálně rozdělí na počet bloků blížící se počtu záznamů, ale pokud údaj obsahuje např. překlep, může dojít k jeho nepropojení či propojení chybnému. (Christen, 2011)

2.2.2.2 Indexace seřazených sousedů

Základní princip je seřazení databáze na základě hodnot blokovacích klíčů a poté sekvenční roztřídění databázi do jednotlivých bloků na základě posunu okna o fixním počtu řádků. Kandidátní páry jsou poté vybírány pouze z jednotlivých bloků. Tento postup je možné implementovat dvěma způsoby:

1) Přístup založený na *seřazených polích (sorted array based)*

   Záznamy jsou seřazeny abecedně a posunem okna jsou poté vytvořeny sekvence záznamů, z nichž budou generovány porovnávací páry. Kandidátní páry jsou následně vygenerovány ze všech záznamů v současném okně (tedy metodou každý s každým) (Hernandez, 1995).

2) Alternativní přístup je založený na *invertovaném indexu* (*inverted index based)*. Namísto umístění hodnot do seřazených polí, tento postup používá invertovaný index podobný tradiční blokaci. Stejně jako u předchozího postupu jsou záznamy seřazeny abecedně, posunem okna se vytvoří sekvence, ale kandidátní páry jsou generovány ze všech párů, které mají korespondující index.

Window positions	BKVs (Surname)	Identifiers
1	Millar	R6
2	Miller	R2
3	Miller	R8
4	Myler	R4
5	Peters	R3
6	Smith	R1
7	Smyth	R5
8	Smyth	R7

Window range	Candidate record pairs
1 – 3	(R6,R2), (R6,R8), (R2,R8)
2 – 4	(R2,R8), (R2,R4), (R8,R4)
3 – 5	(R8,R4), (R8,R3), (R4,R3)
4 – 6	(R4,R3), (R4,R1), (R3,R1)
5 – 7	(R3,R1), (R3,R5), (R1,R5)
6 – 8	(R1,R5), (R1,R7), (R5,R7)

Fig. 3. Example sorted neighbourhood technique based on a sorted array, with BKVs being the surname values from Figure 2 (and the corresponding record identifiers), and a window size $w = 3$.

Window positions	BKVs (Surname)	Identifiers
1	Millar	R6
2	Miller	R2, R8
3	Myler	R4
4	Peters	R3
5	Smith	R1
6	Smyth	R5, R7

Window range	Candidate record pairs
1 – 3	(R6,R2), (R6,R8), (R6,R4), (R2,R8), (R2,R4), (R8,R4)
2 – 4	(R2,R8), (R2,R4), (R2,R3), (R8,R4), (R8,R3), (R4,R3)
3 – 5	(R4,R3), (R4,R1), (R3,R1)
4 – 6	(R3,R1), (R3,R5), (R3,R7), (R1,R5), (R1,R7), (R5,R7)

Fig. 4. Example sorted neighbourhood technique based on an inverted index and with the same BKVs and window size as in Figure 3.

Obr. č. 4: Demonstrace technik indexace seřazení sousedů. BKV je zkratka pro hodnotu blokovacího klíče (block key value). Převzato z Christen(2011)

Efektivitu výše zmíněných přístupů lze zvýšit použitím adaptivní délky okna. Oproti fixní délce, u které hrozí, že nebude schopna pojmout všechny potenciální kandidátní páry, je délka okna určena dynamicky na základě analýzy odlišnosti hodnot za použití metriky měřící přibližnou shodu řetězce. (Yan, 2007; Christen 2006)

2.2.2.3 Indexace pomocí q-gramů

V případě této metody jsou do bloku umístěny hodnoty, které mají nejen stejnou, ale také podobnou hodnotu blokového klíče. Za předpokladu, že tyto hodnoty jsou znakové řetězce, jde o vytvoření variací pro každou hodnotu za použití q-gramů (výřez řetězce o délce q) a jejich umístění do patřičných bloků. Každá hodnota blokovacího klíče je převedena na seznam q-gramů a sub-seznamy kombinací těchto q-gramových seznamů jsou poté vygenerovány na určitou minimální délku, která je určena uživatelem zvolenou prahovou hodnotou. Tyto sub-seznamy jsou následně převedeny zpět na řetězce a použity jako samotné klíče v invertovaném indexu. (Christen, 2011)

Identifiers	BKVs (Surname)	Bigram sub-lists	Index key values
R1	Smith	[sm,mi,it,th], [mi,it,th], [sm,it,th], [sm,mi,th], [sm,mi,it]	**smmiitth**, miitth, smitth, smmith, smmiit
R2	Smithy	[sm,mi,it,th,hy], [mi,it,th,hy], [sm,it,th,hy], [sm,mi,th,hy], [sm,mi,it,hy], [sm,mi,it,th]	smmiitthhy, miitthhy, smitthhy, smmithhy, smmiithy, **smmiitth**
R3	Smithe	[sm,mi,it,th,he], [mi,it,th,he], [sm,it,th,he], [sm,mi,th,he], [sm,mi,it,he], [sm,mi,it,th]	smmiitthhe, miitthhe, smitthhe, smmithhe, smmiithe, **smmiitth**

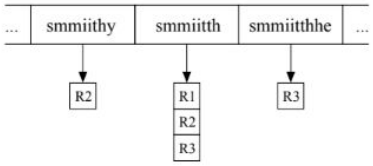


Fig. 5. Q-gram based indexing with surnames used as BKVs, index key values based on bigrams ($q = 2$), and calculated using a threshold set to $t = 0.8$. The right-hand side shows three of the resulting inverted index lists (blocks), with the common BKV highlighted in bold in the index key value column.

Obr. č. 5: Demonstrace metody indexování pomocí q-gramů na základě příjmení použitých jako hodnoty blokovacího klíče. Převzato z Christen(2011)

Dalšími možnými metodami mohou být indexování založené na příponě pole, canopy clustering, nebo indexace založená na řetězcových mapách. (Christen, 2011)

2.2.3 Porovnávání párů záznamů

Kandidátní páry vygenerované indexací vyžadují detailnější porovnání jejich podobnosti. Obecně je podobnost mezi dvěma záznamy určena na základě porovnání několika jejich atributů. V každém případě se nejedná pouze o atribut použitý při indexaci, ale také o další atributy, které jsou dostupné v porovnávaných databázích. Vzhledem k tomu, že různé atributy mohou obsahovat různé typy dat, je zapotřebí použití různých porovnávacích funkcí (Christen, 2008). Pro každý kandidátní pár je porovnáno více kandidátů, výsledkem je vektor obsahující hodnoty podobnosti pro jednotlivé atributy (většinou od 0 do 1, kde 1 je identický pár a 0 je nulová shoda). Tyto vektory γ (viz. kapitola 2.1) jsou posléze použity v procesu klasifikace k rozhodnutí, zda-li se jedná o propojený či nepropojený pár.

2.2.4 Klasifikace párů záznamů

Klasifikace porovnávaných párů záznamů na základě jejich porovnávacích vektorů nebo jejich souhrnné podobnosti je dvou-třídní (binární) nebo tří-třídní klasifikační úloha. V případě binární klasifikace je každý pár buď klasifikován jako *propojený,* nebo *nepropojený*. První třída obsahuje páry záznamů, o kterých se předpokládá, že popisují stejnou entitu, zatímco v rámci druhé třídy naopak páry nepopisují stejnou entitu. Všechny záznamy, které byly odstraněny v procesu indexace, a tudíž nebyly porovnávány, jsou implicitně klasifikovány jako nepropojené.
V případě použití klasifikace do tří tříd může být pár také klasifikován jako *potenciální propojení.* U těchto párů není zřejmé, zda-li reprezentují stejný objekt nebo entitu, a je nutné provést *administrativní přezkum*, tj. pro jejich zařazení do binární klasifikace je nutné, aby záznam přezkoumal odborník.

Klasifikace jako proces je primárně založena na podobnostních hodnotách v porovnávacích vektorech jednotlivých párů. Čím více jsou si záznamy podobné, tím existuje větší pravděpodobnost, že popisují stejnou entitu. Samotný proces může být řízen buď *pod dohledem,* nebo *bez dohledu*. Klasifikace řízená bez dohledu klasifikuje páry nebo skupiny záznamů na základě podobností mezi nimi, aniž by měla přístup k externí informaci, zda-li se doopravdy jedná o stejné nebo odlišné entity. Při klasifikaci pod dohledem má naopak proces přístup k tréninkovým datům, o kterých je známo, zda se jedná (nebo nejedná) o propojené záznamy.

Christen (2012) popisuje několik základních přístupů ke klasifikaci:

1) klasifikace založená na prahové hodnotě,
2) pravděpodobnostní klasifikace,
3) klasifikace založená na nákladech,
4) klasifikace založená na pravidlech.

2.2.4.1 Klasifikace založená na prahové hodnotě

Nejjednodušším způsobem jak klasifikovat páry porovnávaných záznamů je z porovnávacích vektorů vypočítat součet jejich podobnostních hodnot. Hodnota této proměnné (v publikaci označena jako *SimSum*) je poté porovnána s prahovou hodnotou a na základě výsledku je rozhodnuto o propojení páru.

V případě binární klasifikace (propojení a nepropojení) je zapotřebí pouze jedné prahové hodnoty t pro pár záznamů (r_i, r_j):

$$SimSum[r_i, r_j] \geq t \Rightarrow [r_i, r_j] \rightarrow \text{Match},$$
$$SimSum[r_i, r_j] < t \Rightarrow [r_i, r_j] \rightarrow \text{Non-Match}.$$

V případě klasifikace do tří tříd je přidána ještě druhá prahová hodnota. Hodnoty t_l (spodní) a t_u (vrchní) označují prahy, které jsou zapotřebí k přesažení, pro *potenciální propojení* a *propojení.*

$$SimSum[r_i, r_j] \geq t_u \Rightarrow [r_i, r_j] \rightarrow \text{Match},$$
$$t_l < SimSum[r_i, r_j] < t_u \Rightarrow [r_i, r_j] \rightarrow \text{Potential Match},$$
$$SimSum[r_i, r_j] \leq t_l \Rightarrow [r_i, r_j] \rightarrow \text{Non-Match}.$$

2.2.4.2 Pravděpodobnostní klasifikace

Druhým tradičním přístupem ke klasifikaci propojování záznamů je pravděpodobnostní propojování záznamů. V případě absence unikátních identifikátorů entit je nutné použít ostatních atributů dostupných v obou databázích. Vzhledem k tomu, že hodnoty mohou chybět, být špatně uvedené nebo zastaralé, a počet atributů a jejich distribuce nemusí být konstantní, je zapotřebí použít systém vah u jednotlivých atributů. Tyto hodnoty by neměly jenom záležet na obecných charakteristikách atributu, ale také na jeho samotných hodnotách v kandidátních párech. Např. pokud mají oba záznamy hodnotu atributu příjmení "Smith", měla by váha této shody být nižší než např. "Dijkstra", protože příjmení Smith je v USA častější než Dijkstra (Christen, 2012).

Předpokládá se, že záznamy v obou databázích **A** a **B** byly vygenerovány na základě jednoho procesu pro každou z databází. Každý záznam reprezentuje jednoho jedince populace. Záznamy z obou populací jsou vybrány tak, aby existoval překryv (např. někteří pacienti jsou také zákazníci). Pro každého člena je vygenerován záznam s určitými charakteristikami (pro osoby např. jména, datum narození atd.). Zároveň také proces generování záznamů vedl k chybným a chybějícím hodnotám v určitých distribucích. Ve výsledku je možné, že nepropojené entity v množinách **A** a **B** mohou být reprezentovány dvěma záznamy, které mají stejné hodnoty svých atributů, nebo opačně, dva záznamy v množinách **A** a **B** popisující stejnou entitu mohou mít v některých atributech uvedeny odlišné hodnoty.

Při porovnávání párů záznamů je vytvořen porovnávací vektor $\boldsymbol{\gamma}$ pro každý porovnávaný pár. V základní formulaci pravděpodobnostního propojování záznamů se pracuje pouze s binárním porovnáváním, kde hodnota 1 znamená **shodu** a 0 **neshodu** (Fellegi, 1969). Tedy každé $\boldsymbol{\gamma}$ koresponduje s určitým vzorem shody v porovnávacím prostoru $\boldsymbol{\Gamma}$. Pro každý pár porovnávaný počtem **K** porovnávacích funkcí se každé $\boldsymbol{\gamma}$ skládá z vektoru K hodnot vyjadřujících shodu nebo neshodu. Tedy v případě binárního porovnávání bude existovat 2^K různých vzorů.

Pro každý kandidátní pár *r* pravděpodobnostní propojování zvažuje poměr podmíněných pravděpodobností *P*(• | •) :

$$R = \frac{P(\gamma \in \Gamma \mid r \in M)}{P(\gamma \in \Gamma \mid r \in U)}$$

kde **γ** je náhodný vzor shody v porovnávacím prostoru **Γ**. Felegi a Sunter (1969) poté navrhují následující rozhodující pravidlo:

$$\begin{array}{rcll} R \geq t_u & \Rightarrow & r \rightarrow & \text{Match}, \\ t_l < R < t_u & \Rightarrow & r \rightarrow & \text{Potential Match}, \\ R \geq t_l & \Rightarrow & r \rightarrow & \text{Non-match}. \end{array}$$

Prahové hodnoty t_l a t_u jsou určeny z předchozích chybových hranic na základě chybných propojení (příp. chybných nepropojení). (Fellegi, 1969; Herzog, 2007)

Pokud se tedy porovnávací vektor **γ** pro určitý kandidátní pár **r** skládá převážně ze shod, pak poměr R bude vysoký, protože existuje vyšší pravděpodobnost, že **r ∈ M**, než že **r ∈ U**, tedy pár má větší pravděpodobnost, že bude interpretován jako *propojený.* V případě, že je naopak poměr nízký (tedy **r** se skládá převážně z neshod), má **r** větší šanci, že bude interpretován jako *nepropojený*. Zároveň pokud chceme celý proces co nejvíce optimalizovat, měli bychom dbát na to, aby nastavené prahy produkovaly co nejméně potenciálních propojení. (Fellegi, 1969)

Výpočet podmíněných pravděpodobností pro jednotlivé poměry R je podstatný aspekt pravděpodobnostní klasifikace. Tyto pravděpodobnosti jsou podmíněně nezávislé v rámci jednotlivých atributů použitých ve fázi porovnávání párů pro kalkulaci vektoru **γ.** Za tohoto předpokladu, individuální váha shody w_i (kde $1 \leq i \leq K$) může být vypočítána pro každý atribut (nebo pole) *i* na základě m- a u-pravděpodobností:

$$m_i = P([a_i = b_i, a \in A, b \in B] \mid r \in M)$$

pro pravděpodobnost, že oba záznamy mají v porovnávaném atributu stejnou hodnotu a:

$$u_i = P([a_i = b_i, a \in A, b \in B] \mid r \in U$$

pro pravděpodobnost, že záznamy v daném atributu nemají stejnou hodnotu. Kde $\mathbf{a}_i$ a $\mathbf{b}_i$ jsou hodnoty atributu ***i***, které jsou aktuálně porovnávány.

M- a U-pravděpodobnosti jsou také známy jako porovnávací parametry (Herzog, 2007).

2.2.4.3 Klasifikace založená na nákladech

V tradičním způsobu pravděpodobnostního propojování záznamů obě prahové hodnoty $\mathbf{t_l}$ a $\mathbf{t_u}$ jsou nastaveny tak, že celkový počet špatně klasifikovaných kandidátních párů je minimalizován. Při tomto přístupu mohou vzniknout dvě možné chyby: pár záznamů, který popisuje stejnou entitu (oprávněné propojení) je klasifikován jako nepropojený, nebo naopak je oprávněné nepropojení klasifikováno jako propojení. V tradičním modelu se předpokládá, že oba typy chyb mají stejné náklady, nebo jinak řečeno cenu. V některých oborech, např. zdravotnictví, nemusí mít neoprávněné propojení až tak velké následky, jako kdyby záznam propojen nebyl. Klasifikace založená na nákladech umožňuje v tomto procesu tyto chyby váhově rozlišovat. Verykios a kol. (2003) vytvořili nákladově-optimální model založený na Bayesovském přístupu. Vzor shody $\boldsymbol{\gamma}$ je definován v Bayesovském přístupu jako:

$$P(\gamma \in \Gamma \mid r \in M) \geq P(\gamma \in \Gamma \mid r \in U) \Rightarrow r \rightarrow \text{Match},$$

$$P(\gamma \in \Gamma \mid r \in M) < P(\gamma \in \Gamma \mid r \in U)) \Rightarrow r \rightarrow \text{Non-match}.$$

Christen následně na základě Verykiose a kolektivu (2003) vytváří tabulku možných kombinací výsledku procesu klasifikace:

Table 6.1 Costs associated with various matching decisions as proposed by Verykios et al. [263]

Cost	Classification	True match status
$c_{U,M}$	Non-Match	True match (M)
$c_{U,U}$	Non-Match	True non-match (U)
$c_{P,M}$	Potential Match	True match (M)
$c_{P,U}$	Potential Match	True non-match (U)
$c_{M,M}$	Match	True match (M)
$c_{M,U}$	Match	True non-match (U)

Obr. č. 6 Možné kombinace výsledků procesu klasifikace. Převzato z Christen (2011)

Každému z 6 možných řešení lze přiřadit hodnotu C vyjadřující, jak moc je výsledek buď žádaný, nebo naopak nežádaný. Cílem optimalizace tohoto rozhodovacího pravidla je minimalizovat celkové náklady C:

$$C = C_{U,M} \cdot P(r \in \text{Non-match}, r \in M) + C_{U,U} \cdot P(r \in \text{Non-match}, r \in U) + C_{P,M} \cdot P(r \in \text{Potential match}, r \in M) + C_{P,U} \cdot P(r \in \text{Potential match}, r \in U) + C_{M,M} \cdot P(r \in \text{Match}, r \in M) + C_{M,U} \cdot P(r \in \text{Match}, r \in U)$$

kde P(x,y) je společná pravděpodobnost, že pár záznamů **r** byl klasifikován jako třída **x** (kde y ∈ {Nepropojení, potenciální propojení, propojení}), zatímco status opodstatněného propojení **r** je **y** (kde y ∈ {M,U}. Nyní lze aplikovat Bayesovu větu jako náhradu těchto šesti pravděpodobností pravděpodobnostmi toho, že dojde k určitému rozhodnutí za předpokladu, že status opodstatněného rozhodnutí a pravděpodobnosti P(M) a P(U) jsou:

$$P(r = x, r = y) = P(r = x \mid r = y) \cdot P(r = y)$$

kde x a y jsou hodnoty korespondující s výše zmíněnými množinami. Pravděpodobnosti P(r = x | r = y) a P(r = y) mohou být odhadnuty za pomocí trénovacích dat se statusem opodstatněného propojení. (Verykios, 2003)

Christen (2012) dále uvádí, že tento způsob klasifikace není nutně vázaný pouze na pravděpodobnostní ztotožňování záznamů, jak je v dané kapitole prezentováno. Např. v klasifikacích založených na pravidlech (viz násl. kapitola), pravidla mohou být seřazena tak, aby kandidátní páry byly nejprve evaluovány, zda jsou propojené, předtím než se přistoupí k evaluaci, zda se jedná o páry nepropojené, čímž lze dosáhnout také jisté formy prioritizace resp. určování hodnoty jednotlivých typů klasifikace.

2.2.4.4 Klasifikace založená na pravidlech

Tento způsob klasifikace namísto pravděpodobností používá sadu pravidel, pomocí kterých se určuje, zda kandidátní páry jsou propojené, nebo nepropojené (případně detekují potenciální propojení) (Cohen, 2000).

Christen (2012) popisuje jako jádro procesu použití klasifikátoru aplikovaného na podobnostní hodnoty v kombinaci s logickými operátory AND, OR (konjunkce, disjunkce) a NOT (negace). Zpravidla existuje pravidlo P → C, kde P je predikát aplikovaný na podobnostní hodnoty (dostupné v porovnávacím vektoru) pro pár záznamů (r_i, r_j). Predikát P je booleovský výraz v obecném znění:

$$P = (term_{1,1} \vee term_{1,2} \vee ...) \wedge ... \wedge (term_{n,1} \vee term_{n,2} \vee ...)$$

P je napsáno v konjunktivní normální formě jako konjunkce disjunkcí výrazů (Naumann, 2010). Každý výraz je poté aplikován na podobnostní hodnotu jednoho elementu v porovnávacím vektoru γ páru záznamů. Výsledek klasifikace C je následně zařazen do dané klasifikace, pokud je predikát pravdivý, tedy např. je-li hodnota vyšší než určité číslo, záznam je klasifikován jako propojený. Tento systém pravidel může být buď sestaven tak, aby klasifikoval pouze propojení (tedy všechny nepravdivé výroky jsou automaticky klasifikovány jako nepropojené záznamy), nebo může i zahrnovat pravidla pro klasifikaci nepropojených záznamů a potenciálních propojení. Pokud množina pravidel obsahuje pouze pravidla pro klasifikaci propojených záznamů, pořadí pravidel není relevantní. V opačném případě hraje pořadí pravidel velkou roli, protože první pravidlo, u kterého predikát P je vyhodnocen jako pravdivý, je zároveň to pravidlo, které klasifikuje celý porovnávací pár.

2.2.5 Evaluace propojovací kvality a komplexity

Christen (2007, 2012) rozděluje výsledné klasifikace do čtyř kategorií na základě správnosti výsledku:

- ***Pravdivě pozitivní* výsledky** (TP, true positive). Tyto páry záznamů jsou klasifikovány jako propojení a jedná se o opodstatněná propojení. Tedy jedná se o záznamy popisující stejnou entitu.
- ***Nepravdivě pozitivní* výsledky** (FP, false positive). Tyto páry záznamů jsou klasifikovány jako propojené, ale jedná se o neopodstatněné propojení, protože každý ze záznamů se týká odlišné entity.
- ***Pravdivě negativní* výsledky** (TN, true negative). V tomto případě jsou páry klasifikovány jako nepropojené a jejich vztah k reálným entitám tomu odpovídá.
- ***Nepravdivě negativní* výsledky** (FN, false negative). Páry záznamů jsou klasifikovány jako nepropojené, ale ve skutečnosti se jedná o stejnou entitu.

Na základě poměru počtů výskytů těchto výsledků lze určit následující míry indikující kvalitu procesu ztotožňování záznamů:

- **Přesnost (accuracy)**

  $$acc = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|}$$

  Jedná se o poměr správně určených párů proti jejich celkovému počtu. Jedná se o často užívanou míru vhodnou pro vyvážené klasifikační problémy. Nicméně vzhledem k tomu, že tato míra používá i pravdivě negativní výsledky, které v drtivé většině případů převažují, hodnoty přesnosti bývají velmi vysoké i přes chybná přiřazení. Např. chybná klasifikace všech záznamů jako nepropojených může přesto vést k vysoké přesnosti vzhledem k tomu, že velký počet párů, které jsou porovnávány, nejsou propojené. Tedy přesnost není vhodná míra pro ztotožňování záznamů a deduplikaci a neměla by být pro tyto účely používána (Christen, 2007)

- **Preciznost (precision)**

$$prec = \frac{|TP|}{|TP| + |FP|}$$

Tato míra je také někdy označována jako *pozitivní prediktivní hodnota* (Cohen, 1998). Jedná se o poměr opodstatněných propojení vůči všem propojením bez ohledu na jejich správnost. Tato míra se často používá v měření kvality výsledků vyhledávání (Christen, 2007).

- **Výtěžnost (recall, true positive rate)**

$$rec = \frac{|TP|}{|TP| + |FN|}$$

Je podíl pravdivě klasifikovaných pozitivních výsledků vůči všem, které by v ideálním případě propojeny být měly. Společně s precizností se používá v tzv. precision-recall grafech. Podobně jako u preciznosti výtěžnost nezahrnuje počet pravdivě negativních výsledků a lze ji tedy využít i v případech, kdy počty propojených a nepropojených záznamů nejsou vyvážené.

- **F-míra (F-measure)**

$$fmeas = 2 \times (\frac{prec \times rec}{prec + rec})$$

Tato míra kombinuje preciznost a výtěžnost a má vysokou hodnotu pouze v případě, kdy jsou obě hodnoty vysoké. Prakticky se jedná o harmonický průměr dvou hodnot.

- **Specificita (specificity, true negative rate)**

$$spec = \frac{|TN|}{|TN| + |FP|}$$

Podobně jako u přesnosti, tato hodnota by se měla využívat pouze u vyváženého počtu propojených a nepropojených záznamů. V případě, kdy počet nepravdivě pozitivních výsledků je velmi malý oproti počtu pravdivě negativních výsledků (což je velmi pravděpodobné v případě propojování záznamů), bude specificita nakloněna ve prospěch pravdivě negativních výsledků.

- **Míra nepravdivých propojení (false positive rate, fall-out)**

$$fpr = \frac{|FP|}{|TN| + |FP|}$$

Zde je nutno podotknout, že **fpr = (1-spec)**. Opět vzhledem k tomu, že míra počítá s pravdivě negativními výsledky, neměla by se používat při ztotožňování záznamů, stejně jako přesnost a specificita.

Christen (2012) na základě jím citovaných zdrojů uvádí, že nejvíce používanou mírou v počítačové vědě je *přesnost*. Nicméně preciznost, výtěžnost a f-míra zaznamenávají nárůst v popularitě postupně s uvědoměním si jejich nevýhod. (Christen, 2007)

2.3 Výzvy a překážky při ztotožňování záznamů

Integrace dat z různých zdrojů se skládá ze tří úkonů. Prvním je *schema matching*, které se zabývá mapováním databázových tabulek, atributů a konceptuálních struktur z různých databází, které si odpovídají typem obsažené informace. Druhým úkonem je *data matching* - proces identifikace a ztotožňování individuálních záznamů v obou databázích, které se vztahují ke stejné entitě. Speciálním případem je *detekce duplikátů.* Posledním krokem je *data fusion* - proces spojování párů nebo skupin záznamů do jednoho čistého a konzistentního záznamu reprezentujícího entitu. Při aplikaci na jednu databázi se tomuto procesu říká *deduplikace* (Rahm, Do, 2000).

Obr. č. 7: Schéma procesu Integrace dat

Úkol identifikace a ztotožňování záznamů referujících o stejných entitách v jedné, nebo více databázích je náročný z několika důvodů. Čtyři hlavní důvody jsou nedostatek unikátních identifikátorů a kvalita dat, výpočetní náročnost, nedostatek trénovacích dat odpovídající pravé shodě a ochrana osobních údajů a diskrétnost (Christen, 2012).

2.3.1 Nedostatek unikátních identifikátorů a kvalita dat

Databáze, které vyžadují deduplikaci či ztotožňování záznamů, obvykle neobsahují jednoznačné identifikátory či klíče (např. rodné číslo, ISBN, EAN, DOI, ORCID atd.). Ale v případě, že jsou tyto identifikátory dostupné, celý proces se dá zjednodušit na propojení databází za použití SQL příkazů. Nicméně i v těchto případech je potřeba, aby byl pracovník velmi pečlivý a řešení robustní, protože i malé chyby mohou způsobit velké množství špatně propojených záznamů.

2.3.2 Výpočetní náročnost

Pokud chceme zjistit, zda existuje pár záznamů odpovídajících stejné entitě ve dvou databázích, musíme při jejich propojování nejjednodušším možným řešením porovnat každý záznam z jedné databáze se všemi ostatními z té druhé. V případě deduplikace jedné databáze je nutné zase záznam porovnat se všemi ostatními v té samé databázi. Výpočetní náročnost tedy roste s velikostí databáze kvadraticky, zatímco počet potenciálních propojení roste pouze lineárně. Za předpokladu, že neexistují duplikáty, odpovídá počet možných propojení počtu záznamů v menší ze dvou porovnávaných databází.

Hlavní výzvou v oblasti výpočetní náročnosti je efektivní eliminace párů. Tedy snaha o eliminaci těch potenciálních párů, které nemají dostatečně vysokou pravděpodobnost propojení a výběr těch, které tuto šanci mají co nejvyšší.

2.3.3 Nedostatek trénovacích dat odpovídající pravé shodě

V mnoha aplikacích není znám opravdový stav počtu záznamů, které odpovídají stejné entitě. Tedy není možné ověřit, zda opravdu proces byl úspěšný, nebo ne. Tento proces je tímto odlišný od data miningu a machine learningu, kde jsou trénovací data obvykle dostupná. Bez dodatečných informací (jako například dodatečný rozhovor s osobou, ověřující správnost vyplněných údajů) není možné zajistit kontrolu, že jsou opravdu záznamy propojeny správně, což bývá problém zejména u velkých databází.

2.3.4 Ochrana osobních údajů a diskrétnost

Propojování záznamů často spoléhá na použití osobních informací jako jména, adresy, data narození. Soukromí a důvěryhodnost jsou faktory, které je potřeba během procesu vzít v potaz, zejména pokud jsou porovnávány záznamy dvou organizací, nebo výsledky použity externí organizací či individuálními osobami, jako jsou akademičtí pracovníci. Analýza propojených dat má potenciál odhalit aspekty osob nebo skupin, které nejsou v rámci jedné databáze rozpoznatelné. Cílem je tedy podpořit propojování záznamů mezi organizacemi bez toho, aby bylo narušeno soukromí osob, kterých se data týkají.

3. Přibližná shoda znakových řetězců

Úplně přesná shoda znakových řetězců, zejména u delších, bývá často příliš přísné kritérium. V řadě aplikací se osvědčuje hledání shody jen přibližné. V následující kapitole bude popsáno pět nejčastěji používaných metrik vyjadřující přibližnou shodu znakových řetězců:

1) Levenštejnova vzdálenost,
2) Jarova vzdálenost,
3) Jaro-Winklerova vzdálenost,
4) Kosinová vzdálenost q-gramových profilů,
5) Jaccardův koeficient q-gramových profilů.

Tyto metody jsou založeny buď na tzv. *editační vzdálenosti* (tj. minima počtu jistých elementárních operací nutných ke změně jednoho z řetězců na druhý), nebo vzdálenosti založené na porovnání výskytů q-gramů (podřetězců délky q).

3.1 Metody založené na editační vzdálenosti

3.1.1 Levenštejnova vzdálenost

Levenštejnova vzdálenost je minimální počet znaků, které je nutné změnit, přidat, nebo odebrat, aby byly oba řetězce stejné.

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & \text{pokud } min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i,-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} & \text{v ostatnich pripadech} \end{cases}$$

kde je $1_{x \neq y} = 0$ pokud x = y, jinak 1.

Pokud porovnáváme dvě identická slova (např. s =“test” a t =“test”), pak je Levenštejnova vzdálenost lev(s,t) = 0, protože není zapotřebí žádné transformace. Pokud ale t změníme na “text”, lev(s,t) bude rovna 1, protože je zapotřebí jedné substituce. Lze tedy říct, že čím více se od sebe dva řetězce liší, tím je hodnota vyšší (Levenshtein, 1965).

Výpočet lze realizovat následujícím algoritmem, jehož původ není jednoznačný (Navarro, 2001):

Krok	**Popis**
1	Množina **n** má délku **s** Množina **m** má délku **t** Pokud **n = 0**, vrať **m** a ukonči proces Pokud **m = 0**, vrať **n** a ukonči proces Vytvoř matici obsahující řádky 0..m a sloupce 0..n
2	Naplň první řadu hodnotami 0..n Naplň první sloupec hodnotami 0..m
3	Pro všechny znaky **s[i]** (**i** od **1** do **n**)
4	Pro všechny znaky **t[j]** (**j** od **1** do **m)**
5	Pokud **s[i]** je roven **t[j]**, ***cena operace*** je 0 Pokud **s[i]** není roven **t[j]**, ***cena operace*** je 1
6	Nastav element **d[i,j]** matice jako roven minimu z následujících: a) Elementu přímo nad **d[i,j]** plus 1: **d[i-1,j] + 1** b) Elementu přímo nalevo od **d[i,j]** plus 1: **d[i,j-1] + 1** c) Elementu diagonálně nad a doleva od **d[i,j]** plus hodnota z kroku 5: **d[i-1, j-1] +** ***cena operace***
7	Po skončení iterací 3-6 je Levenštejnova vzdálenost zapsána v elementu **d[n,m]**

Postup programu na příkladě transformace slova “PLES” na “PES” by byl následující:

Krok 1:

		P	**L**	**E**	**S**
P					
E					
S					

Nejprve je vytvořena matice, jejíž počet sloupců a řádků odpovídá délkám slov, s přidaným jedním sloupcem a jedním řádkem na začátek. Dále je vyplněn první sloupec a první řádek. Hodnoty v polích označují počet operací, které jsou potřeba k transformaci mezi řetězci.

Krok 2:

V rámci tohoto kroku jsou hodnoty vyplněny bez použití kalkulace počtu transformací. Jedná se o transformaci z prázdného (nebo do prázdného) řetězce a tedy je zapotřebí pouze vkládání. V případě prvního sloupce tedy vytváříme slovo PES z prázdného řetězce, tedy pro jeho celou délku by bylo potřeba přidat 3 písmena (celkem 3 operace). Řetězec PE by vznikl pouze dvěma operacemi atd. Tímto způsobem lze také snadno vyplnit první řádek, který vzniká naopak výpočtem, kolik písmen je třeba odstranit, aby vznikl prázdný řetězec.

		P	L	E	S
	0	1	2	3	4
P	1				
E	2				
S	3				

Kroky 3 až 6:

V případě, kdy ani na začátku, ani na konci transformace nestojí prázdný řetězec, algoritmus už potřebuje i operaci pro zápis, nebo odstranění znaků. Ty si budeme demonstrovat na iteraci i=1, nicméně následně jsou aplikovány na všech dalších operacích stejně.

Při transformaci řetězce P na řetězec P není třeba žádné operace. Program tedy zapíše hodnotu 0 (v programu zapsáno jako pravidlo c z 6. kroku - je použita hodnota elementu diagonálně nad a doleva plus 0).

		P	L	E	S
	0	1	2	3	4
P	1	0			
E	2				
S	3				

Na dalším řádku už transformujeme řetězec “P” na řetězec “PE”. Člověku je jasné, že stačí pouze přidat písmeno E. Nicméně program k tomu musí dojít složitější cestou, a to tak, že porovná tři hodnoty ze sousedních polí (viz krok 6) a vybere tu nejnižší. Tato hodnota reprezentuje nejmenší možný počet transformačních operací nutných k dosažení cílového řetězce.

Program tedy zvažuje mezi:

a) Přidáním znaku E (hodnota elementu přímo nad plus 1, tedy 0+1) → pouze 1 operace
b) Odebráním prvního znaku a poté přidáním dvou znaků (hodnota elementu přímo nalevo plus 1, tedy 2+1) → celkem 3 operace
c) Transformací 1 znaku a přidáním druhého znaku (hodnota elementu diagonálně nad a doleva plus hodnota z kroku 5, tedy 1+1) → celkem 2 operace.

Minimum možných operací je tedy **1** adice a tedy je do matice zapsáno 1.

		P	**L**	**E**	**S**
	0	**1**	**2**	**3**	**4**
P	**1**	**0**			
E	**2**	**1**			
S	**3**				

Podobným způsobem je postupně zaplněna celá matice:

		P	**L**	**E**	**S**
	0	**1**	**2**	**3**	**4**
P	**1**	**0**	**1**	**2**	**3**
E	**2**	**1**	**1**	**1**	**2**
S	**3**	**2**	**2**	**2**	**1**

V poslední operaci nakonec provádíme skutečnou transformaci řetězce PLES na PES a vyjde nám, že Levenštejnova vzdálenost pro daný příklad je **1**.

3.1.3 Jarova vzdálenost

Tato metrika měří podobnost mezi dvěma řetězci v rozmezí 0 až 1, kde 1 symbolizuje naprosto identické řetězce. Je definována jako 1 – sim_j, kde sim_j je Jarova podobnost (Jaro, 1989) dvou řetězců s_1 a s_2. Ta je definována vztahem:

$$sim_j = \begin{cases} 0 & \text{pokud } m = 0 \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}) & \text{v ostatnich pripadech} \end{cases}$$

kde:

- $|s_i|$ je délka řetězce *i*
- *m* je počet shodných znaků
- *t* je polovina počtu transpozic

Dva znaky z s_1 a s_2 jsou považovány za shodné, pouze pokud jsou stejné a ne dále než:

$$\lfloor \frac{max(|s_1|, |s_2|)}{2} \rfloor - 1$$

Každý znak z s_1 je porovnán se všemi shodnými znaky v s_2. Počet transpozic t je definován jako polovina počtu shodných znaků, které v obou řetězcích jsou na jiných místech.
Například porovnáváme-li řetězce BOUDA a HROUDA:
m = 4
s1 = 5
s2 = 6

$$[\frac{max(|s_1|, |s_2|)}{2}] - 1 = 2$$

t = 4

$$sim_j = \frac{1}{3}(\frac{4}{5} + \frac{4}{6} + \frac{4-2}{4}) \approx 0.65$$

3.1.4 Jaro-Winklerova vzdálenost

Jaro-Winklerova vzdálenost je varianta Jarovy vzdálenosti navržená Wiliamem E. Winklerem v roce 1990. Používá dodatečný člen, který vzdálenost zmenšuje, pokud se řetězce shodují ve svém prefixu (prvních několika znacích).

Jaro-Winklerova podobnost (Winkler, 1990) modifikuje Jarovu podobnost:

$$sim_w = sim_j + lp(1 - sim_j)$$

kde:

- sim_j je Jarova podobnost pro řetězce s_1 a s_2 ,
- l je délka společného prefixu na začátku řetězce s maximem 4 znaků,
- p je konstantní škálující faktor určující zvýšení podobnosti, pokud jsou prefixy shodné. Nesmí být vyšší než ¼. Standardně se pracuje s hodnotou p=0,1.

Samotnou Jaro-Winklerovu vzdálenost pak získáme výpočtem:

$$d_\omega = 1 - sim_\omega$$

3.2 Metody založené na vzdálenosti *q*-gramů

Q-gram je řetězec skládající se z q po sobě jdoucích znaků. Q-gramy spojené s řetězcem s jsou získány postupným posunem okna o q znacích přes řetězec s a registrací vyskytujících se podřetězců. Například digramy (2-gramy) v řetězci “foo” jsou “fo” a “oo”. Samozřejmě, tato procedura selže, jestliže $q > |s|$ nebo $q = 0$. (Van Der Loo, 2014)
Uvedeme dvě metody, které pro stanovení míry shody řetězců používají q-gramy konstruované z porovnávaných řetězců.

3.2.1 Jaccardův koeficient

Jaccardův koeficient měří podobnost dvou množin zjištěním, kteří členové jsou v obou množinách a kteří ne. Výsledná hodnota je udávána v procentech. (Jaccard, 1912)

Koeficient měří podobnost mezi konečnými soubory vzorků a je definován jako velikost průniku dělená velikostí sjednocení souborů vzorků:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Odečtením Jaccardova koeficientu od jednotky se získá Jaccardova vzdálenost. Ta je mírou odlišnosti dvou množin. Jaccardovu vzdálenost můžeme ekvivalentně získat jako rozdíl velikostí sjednocení a průniku dvou množin děleno velikostí sjednocení množin:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

Pokud $Q(s;q)$ značí množinu unikátních q-gramů vyskytujících se v textovém řetězci s, lze Jaccardovu vzdálenost řetězců s a t vyjádřit následovně:

$$d_{jaccard}(s,t;q) = 1 - \frac{|Q(s;q) \cap Q(t;q)|}{|Q(s;q) \cup Q(t;q)|}$$

Svislé čáry ($|\cdot|$) zde značí počet prvků množiny.

Jaccardova vzdálenost nabývá hodnot od 0 do 1, přičemž 0 se nabývá právě jen při úplné shodě obou řetězců, tedy pro $Q(s;q) = Q(t;q)$. (Van Der Loo, 2014)

3.2.2 Kosinová vzdálenost *q*-gramů

Kosinová vzdálenost měří kosinus úhlu svíraného mezi dvěma vektory v promítnutém multidimenzionálním prostoru všech možných q-gramů.

Označme Σ množinu všech přípustných znaků v řetězcích. Označme $v(s;q)$ vektor dimenze $|\Sigma|^q$, jehož koeficienty reprezentují počet výskytů všech možných q-gramů v řetězci s. Jakmile máme definovány tyto vektory, můžeme definovat míru jejich podobnosti jako skalární součin normovaný součinem jejich velikostí. V geometrické interpretaci uvedených vektorů se takovýto normovaný skalární součin představuje jako kosinus úhlu těmito vektory svíraného. Vzdálenost opět zavedeme jako rozdíl jednotky a míry podobnosti:

$$d_{cos}(s,t;q) = 1 - \frac{v(s;q) \cdot v(t;q)}{||v(s;q)||_2 ||v(t;q)||_2}$$

$\| v \|_2$ zde indikují standardní euklidovskou normu, tj. $\sqrt{{v_1}^2 + {v_2}^2 + \ ... \ + {v_n}^2}$, kde *n* značí dimenzi vektoru *v*.

Kosinová vzdálenost pro *s* = *t* se rovná 0.

Rovná se 1, jestliže *s* a *t* neobsahují žádné společné *q*-gramy. (Van Der Loo, 2014)

4. Specifika metadat vědeckých publikací

Vědeckými publikacemi rozumíme literaturu, jež přináší nejnovější vědecké poznatky z určitého vědního oboru či disciplíny nebo shrnuje výsledky dosavadního vědeckého bádání a je určena kvalifikovaným čtenářům (Matušík, 2003).

V procesu vědecké komunikace hrají publikace důležitou roli, protože slouží jako nosiče informací. Pro tyto publikace, stejně jako pro ostatní informační zdroje obecně, platí, že v případě efektivního uložení a kvalitního dlouhodobého uchovávání je snazší dané zdroje opětovně nalézt při vyhledávání, což je pro vědu a výzkum velmi důležité, zejména během získávání informací pro navazující výzkum, ověřování stávajících informací nebo vytváření nových poznatků. Aby byly tyto podmínky splněny, je potřeba dokumenty jednoznačně identifikovat, aby nedošlo k jejich záměně.

Pro vědecké publikace existuje mnoho různých metadatových popisů, podle toho v jakém zdroji je informace prezentována. To může být přímo ve vědecké literatuře samotné (formou odkazu na literaturu), nebo pak v bibliografických a citačních databázích nejrůznějšího zaměření, online repozitářích vědecké literatury a v informačních systémech o aktuálním výzkumu.

4.1 Specifika zápisu metadat

V případě metadat vědeckých publikací platí podobné konvence pro jednotlivá běžná pole jako u ostatních typů publikací či údajů. Většina dat obsažená v databázi jsou textové řetězce, ale i u nich lze vnímat určité odchylky podle typu údaje, jež pole obsahuje.

4.1.1 Názvy publikací

Název publikace většinou shrnuje obsah do několika slov, příp. věty. Jeho sekundární funkcí může být zachytit čtenářovu pozornost a diferencovat publikaci od ostatních v daném oboru. Název publikace by měl být jednoduchý, jednoznačný a samozřejmý. Narozdíl od ostatních metadat, pro názvy vědeckých publikací existují různá doporučení, co se jejich obsahu týče, ale existuje jenom minimum konvencí, které by se týkaly jejich formátu. V databázích je tento údaj

většinou uložen buď v jednom poli, případně je rozdělen na hlavní název a podnázev. Jedná se o relativně krátké řetězce, vhodné jak k exaktnímu, tak přibližnému porovnávání. Problémy mohou nastat pokud neexistují konvence pro zápis určitých údajů a na základě toho vznikají nepřesnosti. Typickými příklady může být např. zápis zkratek či chemických vzorců nebo jiných údajů, jež vyžadují specifické formátování. V případě, kdy se porovnávají dva záznamy identických vědeckých publikací, ale jedna má v názvu některá slova zkrácenou formou a druhá v plné délce, může dojít k nepropojení. V případě chemických vzorců může např. dojít k významné změně v obsahu názvu, pokud se změní jen jedno písmeno či číslo ve vzorci. Dále lze také zmínit jako problematické datumy, např. zápis s jinou formou interpunkce může pro algoritmus působit jako velká změna, to samé se může stát, i pokud je přikládána velká váha pořadí znaků a existují rozdíly v pořadí údajů v datumu (např. rok-měsíc-den vs. den-měsíc-rok). Některé názvy publikací také mohou obsahovat údaje, jež se mohou jevit jako gramatické chyby, ale ve skutečnosti je se jedná o zápis např. značky, jiného slova, či zkratky (např. název populární taxi služby Lyft je při přibližné shodě řetězců velice podobný anglickému slovu “lift”). Velké rozdíly může u publikací z oboru molekulární genetiky také způsobit záměna znaků v zápisu řetězců DNA za pomocí písmen A, C, T, G.

4.1.2 Abstrakty

Abstrakty jsou podobně jako názvy publikací uloženy ve formě dlouhých řetězců. Platí pro ně stejné problémy jako pro názvy publikací, ale projevují se v menší míře, protože objem ostatního textu je větší. Další výhodou abstraktů je, že se většinou nepublikují ve více verzích a tím pádem jsou méně různorodé napříč různými identickými objekty s výjimkou chyb způsobených překlepy či jiným formátováním. Nevýhodou naopak může být, že ne vždy jsou abstrakty v databázi obsaženy přímo, ale pouze jako externí reference a jejich zajištění tak může být obtížnější oproti jiným údajům. Zároveň jejich zpracování je více výpočetně náročné a může trvat dlouhou dobu.

4.1.3 Autoři, spoluautoři a jiné osoby spojené s vědeckou publikací

V metadatech publikace jsou uváděny osoby, jež jsou s ní nějakým způsobem spojeny. Kromě samotných metadat osob je také možné využít jednoznačný identifikátor osoby jako ORCID, nebo ResearcherID. Výčet možných rolí, jež osoby mohou v rámci vědecké publikace plnit, je dlouhý. U většiny publikací se minimálně uvádí autor a spoluautoři, záznamy ale mohou obsahovat editory, autory ilustrací, předmluv či osoby, jež byly během práce konzultované. V případě, kdy existují rozdíly mezi způsobem, jakým jsou tyto osoby v metadatech evidovány (např. jedna z databází eviduje všechny autory v jednom poli, zatímco druhá má jedno pole pro hlavního autora a druhé pole pro spoluautory), může dojít ke snížení pravděpodobnosti, že dojde k propojení záznamů. Dalším problémem může být tzv. "hyperautorství". Jedná se o stav, kdy článek je prezentován jako dílo obrovského množství autorů[1]. Mimo nadměrné množství či absenci osob spojených s publikací je také třeba rozlišovat způsoby zápisu jejich jmen, příp. dalších dostupných údajů. Je potřeba zohlednit regionální specifika jmen, jež se nemusí v databázovém řešení projevit. Např. nizozemská příjmení často používají "van" jako předponu (holandské *voorvoegsel*), ale zároveň existuje i příjmení Van. Tato předpona nemusí být formou zápisu rozlišena a může tak dojít k záměně. Dalším příkladem mohou být islandská jména, která se skládají z rodného a patronymického či matronymického jména a nejsou dědičná, takto může poté dojít k záměně dvou osob, protože mají stejné rodné jméno a jméno otce. Další možný problém při ztotožňování je formát zápisu těchto jmen v samotné databázi, pokud jedna databáze eviduje jména formou [jméno, příjmení] a druhá [příjmení, jméno] a není to zohledněno v procesu přípravy a čištění dat, může dojít k nepropojení identických záznamů. Poslední, veskrze minimálně se objevující raritou jsou případy, kdy spoluautorem publikace není člověk, ale zvíře (Erren, 2017). Tato možnost je vědeckou komunitou víceméně tolerována, problém ale může nastat v případě, kdy je toto autorství poté žertem použito i v jiných publikacích (např. F.D.C. Willard nebo H.A.M.S. ter Tisha).

[1] Typické příklady jsou následující články:
https://doi.org/10.1103/PhysRevLett.114.191803 a https://doi.org/10.1534/g3.114.015966

4.1.4 Forma

V akademické sféře jsou nejčastějšími formami publikace online nebo v tištěné podobě. Nicméně někteří dodavatelé stále nabízí články na CD nebo mikrofiších.. Je ale zapotřebí při ztotožňování záznamů mezi nimi rozlišovat. Vědecké publikace často vychází v ve více formách, musíme tedy rozlišovat zda v dané situaci sledujeme jednotlivá zhmotnění díla nebo naopak pouze jeho vyjádření. V případě, kdy jsou dobře zapsány ostatní metadata bez zohlednění nosiče, mohlo by dojít k propojení dvou odlišných objektů, přestože to není cílem.

4.1.5 Vydání

Vydání označuje určitou verzi publikovaného textu. V akademické sféře se toto verzování zpravidla používá u obsáhlejších publikací, kdy je potřeba text aktualizovat o nové poznatky či opravit chyby. U akademických článků se často rozlišuje mezi tzv. preprintovou verzí (podoba článku před peer review, editací a přijetím do odborného časopisu), postprintovou verzí (přijato do odborného časopisu se všemi změnami navrhnutými v procesu peer review) a finální publikovanou verzí (verze přijatá k publikaci obohacená vydavatelem o klíčová slova, prostředky pro indexaci apod.). Stejně jako u typu nosiče je mezi jednotlivými vydáními potřeba rozlišovat, zda je cílem sledovat jednotlivá zhmotnění díla. Pokud naopak cílem je sledovat jen vyjádření, často se použijí údaje toho nejčerstvějšího zhotnění, které je pořizovateli metadatového záznamu známo.

4.1.6 Nakladatelské údaje

Nakladatelskými údaji rozumíme informace o nakladateli (vydavateli), datum a místo publikování. V akademické sféře existuje mnoho typů vydavatelů, od velkých nakladatelství po soukromé osoby. U samotného jména a jiných údajů o nakladateli se mohou vyskytovat podobné problémy jako u autorů a názvů publikací, s tím rozdílem, že se jedná o instituce. Tedy mohou vzniknout problémy, pokud je nakladatel uveden se zkratkami a spol. nebo s.r.o nebo pokud názvy obsahují slova jako “nakladatelství”, jež mohou být v některých případech součást oficiálního názvu, a jindy pouze holé konstatování že se jedná o nakladatelství. Dalším možným problémem může být podobně jako u fyzických osob změna jména, kdy jeden zdroj obsahuje

neaktualizovaný údaj, zatímco jiný zdroj už údaj aktualizoval. Otázkou poté je, zda záznam vůbec aktualizovat nebo ne, metadatový záznam publikace má reprezentovat publikaci v některém jejím význačném okamžiku (nejčastěji okamžiku vydání). Pozdější změny většinou po publikaci se nereflektují.

4.1.7 Časové údaje a datumy

Jednotlivé informační zdroje mají různé způsoby formátování časových údajů, nicméně pokud je formát konzistentní, lze mezi nimi velmi snadno v procesu čištění dat převádět. Problémy mohou nastat v případech, kdy jsou mezi oběma porovnávanými záznamy rozdíly v definici časového údaje. Databáze mohou např. mít rozdíly v datech vydání, pokud se liší metoda, kterou určují, kdy je dokument publikován. Dále se jeví jako možná překážka rozdíly v granularitě datumů. Pokud je například narychlo vydáno stejnojmenné vydání s korekcemi o několik měsíců později a tato informace je uvedena pouze v názvu publikace, může dojít k propojení či nepropojení dvou odlišných publikací za předpokladu, že jsou porovnávány názvy a data vydání s granularitou na roky.

4.2 Identifikátory

Jednotlivé údaje o publikacích však většinou neumožňují jednoznačně identifikovat dokument. V některých případech může dojít k záměně dokumentu i v případě kompletní citace. Od 70. let 20. století byly používány bibliografické identifikátory jako ISBN či ISSN, které umožňují jednoznačnou identifikaci. S rozvojem webu přibyly další identifikátory jako URN, nebo DOI (Hakala, 2010).

4.2.1 ISBN

Systém standardního čísla knihy SBN vznikl v roce 1968. ISBN, tedy mezinárodní standardní číslo knihy (international standard book number) bylo zřízeno v roce 1970 mezinárodní agenturou se stejným jménem. V jednotlivých zemích fungují tzv. národní / skupinové agentury ISBN, jež přidělují vydavatelům prefixy identifikátorů a bloky čísel ISBN s vypočtenými kontrolními číslicemi. Národní agentury také budují a spravují databáze čísel ISBN a příslušných

metadat. V České republice plní funkci národní agentury Národní knihovna ČR. Samotný identifikátor má následující strukturu (ISBN Users' Manual, 2017):

- prefix - trojciferná číslice poskytovaná společností EAN international. V současné době jsou používány pouze 978 a 979
- identifikátor skupiny - identifikuje zemi nebo geografickou či jazykovou oblast. Přiděluje Mezinárodní agentura ISBN. Pro ČR a Slovensko se používá 80.
- identifikátor vydavatele - přidělován národní agenturou ISBN, určuje vydavatele publikace
- identifikátor titulu - jednoznačně určuje titul, přiděluje jej vydavatel.
- Kontrolní číslice - jednociferné číslo umožňující kontrolu správnosti zadání ISBN. Vypočítává se na základě algoritmu z předchozích částí identifikátoru.

Počty číslic identifikátorů skupiny, vydavatele a titulu mohou být proměnné, celkem však vždy jde o devět číslic.

4.2.2 ISSN

ISSN (International standard serial number) umožňuje jednoznačnou identifikaci seriálově pokračujících publikací (periodika). Systém mezinárodní sítě se skládá z mezinárodního centra ISSN, jež slouží k řízení a koordinaci a udržuje centrální registr. Dále jsou zřízena národní centra ISSN, která vytváří bibliografická data a mají na starosti přidělování identifikátorů na základě žádosti vydavatele. V České republice plní tuto funkci Národní technická knihovna. Identifikátor má podobnou skladbu jako u ISBN. Celkem se skládá z 8 znaků, kde první 4 znaky označují číselný kód bloku podle centra, jež kód přiřazuje periodiku, následující 3 znaky pak představují pořadové číslo periodika a poslední je kontrolní znak.

4.2.3 URN

URN (Uniform Resource Name) vznikl v roce 1994 zveřejněním internetového standardu RFC1737. Identifikátor lze snadno rozšířit a škálovat na nové zdroje, navíc jej lze integrovat i do jiných, již existujících schémat. Identifikátor se skládá ze schéma, identifikátoru jmenného prostoru a specifického řetězce jmenného prostoru. Schéma je vždy označeno "urn:". Identifikátor jmenného prostoru se skládá z písmen, číslic a pomlček a samotný specifický

řetězec jmenného prostoru poté může obsahovat libovolné znaky. Nejznámějším případem URN je URN:NBN, který je určen pro systémy národních bibliografií jako lokalizační nástroj. V České republice lze uvést jako příklad využití tohoto identifikátoru systém pro trvalou identifikaci a lokalizaci dokumentů českého digitálního kulturního dědictví CZIDLO.

4.2.4 DOI

Zkratka DOI (digital object identifier) znamená digitální identifikátor objektu. Systém vznikl v roce 1997 společnou iniciativou tří obchodních asociací v nakladatelském průmyslu: *Mezinárodní asociace nakladatelů*, *Mezinárodní asociace vědeckých, technických a zdravotnických vydavatelů* a *Americkou asociací nakladatelů*. Následkem vznikla *Mezinárodní asociace DOI* (International DOI Foundation - IDF).

Celý systém je navržen tak, aby fungoval v prostředí Internetu. Identifikátor DOI je permanentně přiřazen objektu, čímž je umožněno poskytnutí persistentního síťového odkazu na aktuální informace o objektu, vč. informace o jeho umístění. Zatím co se informace o objektu (včetně jeho umístění) mohou změnit, identifikátor zůstává pořád stejný. Systém DOI také umožňuje konstrukci automatizovaných služeb a transakcí. Je tak možné systém aplikovat na management informací, dokumentaci lokace, přístupu a persistentní identifikaci jakékoliv formy dat. Obsah objektu asociovaného s identifikátorem je popsán DOI metadaty, která jsou založena na strukturovaném rozšiřitelném datovém modelu, kteryúmožňuje popis na jakékoliv úrovni úplnosti a granularity. (ISO 26324, 2012)

Syntaxe DOI se skládá ze dvou částí oddělených lomítkem - prefixem a sufixem. Prefix označuje registrátora identifikátoru a sufix je unikátní řetězec znaků jím zvolený. Prefix má nejčastěji podobu 10.XXXX, kde XXXX je čtyřciferné číslo větší nebo rovno 1000. Číslo 10 označuje jmenný prostor DOI a odlišuje jej od ostatních identifikátorů v systému Handle. Prefix může být také dále rozdělen tečkami, např. 10.XX.X.X (DOI Handbook, 2014). Sufix je řetězec znaků libovolné délky a je unikátní v rámci prefixu, je tedy možné, aby dva registrátoři přiřadili stejný suffix, nicméně kódy budou stále unikátní vzhledem k odlišným údajům v prefixu. V konstrukci sufixu někteří vydavatelé preferují sekvenční řazení, zatímco jiní využívají identifikátor, jenž byl

předtím přiřazený jinou autoritou. DOI tak může mít formu např. 10.1007/978-3-319-68619-6_60 nebo 10.1088/1742-6596/1056/1/012002.

Přiřazování DOI se řídí několika principy((DOI Handbook, 2014):

- DOI by nemělo být použito jako náhrada jiných ISO identifikačních schémat
- DOI může být přiřazeno jakémukoliv objektu v kterékoliv situaci, kdy je potřeba jej odlišit od ostatních objektů
- DOI je stavěno jako "digitální identifikátor objektu", ne "identifikátor digitálního objektu"

Dále pro přiřazování existuje několik dalších pravidel:

- **Granularita**

  DOI může být přiřazeno jakémukoliv objektu, bez ohledu na to, v jakém rozsahu je součástí jiné větší entity. DOI identifikátory mohou být přiřazeny v jakékoliv míře přesnosti a granularity, jež registrátor uzná za vhodnou. Např. v textových materiálech mohou být zvláštní DOI přiřazeny jak publikaci jako dílu, tak specifickému vydání této publikace nebo určité kapitole, straně či ilustraci.
- **Změna objektu**

  Upravenému objektu může být přiřazeno nové DOI v tom případě, kdy je nutnost rozlišit novou verzi od originálu. Mezinárodní asociace DOI v daném směru nestanovuje zvláštní pravidla, nicméně jednotlivé agentury mohou stanovit vlastní opatření specifická pro jejich komunitu.

- **Popis**

  Přiřazení DOI vyžaduje, aby registrátor poskytl metadata popisující objekt, jemuž je identifikátor přiřazován. Metadata objektu jej popisují na úrovni, jež umožňuje jeho rozlišení jakožto unikátní entity v rámci DOI systému.
- **Unikátnost**

  Každé DOI specifikuje pouze jeden jediný objekt v rámci systému.

- **Trvalost**

  Existence identifikátoru by neměla být časově limitována. Podoba identifikátoru se nemění při změně vlastníka práv či správce. DOI systém poskytuje prostředky na zajištění pokračování interoperability skrze výměnu informací o identifikovaných entitách (v minimálním rozsahu DOI a popis objektu)

V registraci DOI vědeckých publikací má zcela nejvýznamnější postavení registrační agentura Crossref. Další registrační agentury pokrývají specifické regiony či zvláštní případy použití.

4.3 Hodnocení kvality metadat

Procesu propojování záznamů předchází zhodnocení kvality metadat. Posuzování kvality nelze provádět bipolárně - tedy formou posudku, zda-li je kvalita dostatečná nebo ne. Je zapotřebí více pragmatického a manažerského náhledu na problematiku. Realistické přístupy balancují funkcionalitu metadat proti platným omezením. Tím je zaručeno, že metadata poskytují maximální přínos a šetří zdroje vynaložené na jejich tvorbu, organizaci a kontrolu. Mezi obecně rozpoznávané indikátory kvality patří: (Bruce, 2004)

- **Celistvost**

  Data by měla být celistvá ve dvou smyslech. Buď se jedná o kompletnost v rámci rozsahu popisu, nebo o kompletnost využití definovaných metadat. Metadata by měla objekt popisovat v plném rozsahu a co nejúsporněji. Je možné si téměř vždy představit vyšší úroveň detailnosti popisu, ale náklady nutné na přípravu a údržbu detailnějších informací nemusí být vyváženy předpokládaným přínosem. Dále by měla všechna definovaná metadata být na jednotlivé objekty použita v co nejvyšší možné míře.

  Není dobré definovat konkrétní sadu prvků, pokud většina prvků není nikdy použita, nebo pokud se na jejich použití nelze spolehnout v celé kolekci.

- **Přesnost**

  Metadata by měla být přesná v popisu objektů jež reprezentují. V minimálním rozsahu by se mělo jednat o korektnost a faktickou správnost. Přesnost se ale také vztahuje na minimalizaci typografických chyb, standardizaci jmen osob a zkratek.

- **Provenience**

Provenience metadat je častým základem pro hodnocení kvality. Jedná se o informace, jež poskytují náhled na identitu, expertízu a zkušenost osoby, která metadata vytvářela nebo připravovala. Je samozřejmě možné také použít dobře známé, či certifikované metodologie jako garanty důvěryhodnosti a kvality. Vědci a statistici jsou obvykle zdatní v posuzování kvality dat na základě metod použitých při jejich tvorbě a manipulaci, zejména v situacích, kdy nelze ověřit objekty jednotlivě.

- **Míra shody s očekáváním**

  Standardní metadatové elementy a profily aplikací, jež je používají, lze považovat za určitou formu příslibu poskytovatele směrem k uživateli. To znamená, že komunita, která používá tyto systémy, má určité představy a očekávání, týkající se jejich použití. Sady elementů metadatového standardu by měly, obecně vzato, obsahovat takové elementy, jež by cílová skupina uživatelů plánovala použít. Naopak by neměly být obsaženy elementy, které nebudou pravděpodobně použity, nebo jsou nadbytečné, irelevantní, či je nemožné je implementovat. Výběr metadat by měl také odrážet myšlení a představy cílového uživatele o nutných kompromisech v implementaci. Je ale také nutno poznamenat, že je jen velmi zřídka možné, aby poskytovatel metadat naplnil všechna očekávání. Nicméně je zapotřebí očekávání komunity získávat, zvažovat a spravovat realisticky s velkou opatrností.

- **Logická konzistence a koherence**

  Konzistence a koherence se většinou vyskytují v heterogenních, federovaných sbírkách, nebo u sbírkek informací, jež jsou určitým způsobem verzovány. Ve skutečnosti existuje jenom velmi malé množství sbírek v kompletní izolaci, včetně behem jejich vzniku. Téměř vždy existuje potřeba zajištění, že jednotlivé elementy metadat jsou vnímány způsobem, který je konzistentní s definicemi použitého standardu, a jsou tak i prezentovány koncovému uživateli.

- **Včasnost**

  Pro popis včasnosti metadat se používají dva termíny: aktuálnost a zpoždění. Problemy s aktuálností nastanou v případech, kdy se cílový objekt změní, ale metadata zůstanou stejná. Ke zpoždění dochází, když je cílový objekt šířen v době, kdy nejsou všechna metadata dostupná, nebo zjistitelná.

- **Aktuálnost**

  Typickým příkladem chyby aktuálnosti metadat může být např. zastaralý URI identifikátor, jež dávno není pro daný objekt platný. Nicméně skoro jakýkoliv element nebo hodnota se může v průběhu času oddělit od původního objektu nebo cíle, s jakým byla použita. Informační objekty se mohou pohybovat, ať už po policích, webových stránkách nebo konceptuálních mapách v oboru. Metadata s časem ztrácejí kvalitu, pokud nejsou synchronizována.

- **Zpoždění**

  Šíření metadat není nutně spojeno s šířením objektu, jehož se týkají. Nové objekty vyžadují čas na jejich popsání, katalogizaci a kategorizaci, zejména v případech, kdy je zapotřebí lidské síly. Může tak dojít k situaci, kdy objekt musí být rychle šířen a metadata zaostávají pozadu.

- **Přístupnost**

  Metadata, jež nejsou dostupná nebo je nelze pochopit uživatelem, mají nulovou hodnotu. Překážky mohou být jak fyzické, tak intelektuální. Fyzické bariéry se mohou projevovat různými způsoby. Metadata nemusí být přímo asociovány s cílovými objekty, např. kvůli jejich fyzickému oddělení, nebo pokud pocházejí z jiného zdroje. Dalším případem je chybné nebo špatně realizované propojení metadat a objektu, který popisují, z technických důvodů jako zastaralý formát či software, nebo jsou přístupná pouze pro uživatele s předplatným. Metadata i objekty jsou často používány vícero skupinami a způsob jejich šíření se obtížně předpovídá. Poskytovatelé metadat mají zřídka kontrolu nad tím, zda uživatel přistupující k metadatům má dostatečné znalosti k jejich pochopení. Nicméně některé intelektuální bariéry mohou být sníženy pečlivým zvážením potenciálních cílových skupin během návrhu a dokumentace implementace metadat jako takových.

Lze uvést několik výzkumů, jež se zabývaly výskytem těchto chyb v praxi:
Jak databáze WoS, tak Scopus vykazují nepřesnosti dat citací (Van Eck, 2019) jako:

- chybějící reference (záznam, na který je odkazováno, chybí kompletně),
- chyby v referencích (např. špatně uvedené datum publikace),

- chybné reference (odkazování na chybný dokument, který je sice identický názvem autora a např. datem vydání, ale zbytek záznamu obsahuje chybné informace),
- duplicity.

Problémem duplicit se zabývá Valderrama-Zurián (2015). V daném výzkumu zmiňuje několik příčin duplicit v databázi Scopus:

- články publikované v odborném časopise jsou z důvodu chybného mapování napojeny na dva různé časopisy od stejného vydavatele,
- duplicity vzniklé z důvodu změny názvu odborného časopisu,
- duplicity vzniklé rozdíly v zápisu názvů časopisů (včetně tak malých rozdílů jako je pouhá velikost písmen, např. BMJ Case Reports vs. BMJ case reports).

Vědecké publikace jsou často klasifikovány podle typu dokumentu, jako např. výzkumné články, přehledové články či knihy apod. Při analýze přiřazených typů dokumentů na náhodném vzorku publikací indexovaných Web of Science oproti nezávisle přiřazeným typům byla přesnost 94% (Donner, 2017).

Další nekonsistence lze najít v datech publikování dokumentů. Kvůli akceleraci vědecké komunikace díky digitálním technologiím nemusí být pouze rok publikace dokumentu nadále dostačující. V případě použití přesnějších dat publikování zároveň ale dochází k velkým rozdílům v těchto údajích mezi jednotlivými zpracovateli metadatových záznamů. Jednotlivé databáze mají zároveň různé metody indexace dat vydání, jako např. online datum poskytnuté vydavatelem, měsíc, ve kterém bylo dané číslo časopisu vydáno, datum indexace Web of Science nebo i datum první zmínky publikace. Tyto metody nemají v současné době dostatečnou transparentnost a standardizaci. (Haustein, 2015)

4.4 Korekce metadat vědeckých publikací

Metadata vědeckých publikací mohou obsahovat chyby. V kontextu digitálních knihoven můžeme rozlišovat tři různé úhly pohledu na kvalitu dat (Beall, 2006):

- **Absolutní kvalita dat**

  Celková úroveň kvality dat jak digitálního objektu tak jeho metadat.

- **Věrohodná reprodukce kvality dat**

  Kvalita objektů, jež vznikly jinde než v prostředí digitální knihovny. Věrohodná reprodukce znamená, že digitální objekty v repozitáři přesně odpovídají dokumentu nebo objektu v jeho originální formě.
- **Kvalita digital born dat**

  Kvalita dat v digitální knihovně, které vznikly přímo v rámci digitální knihovny. Jedná se o metadata, jež byla vytvořena přímo digitální knihovnou.

Chyby v kvalitě dat můžeme rozdělit na několik kategorií (Beall, 2006). Tyto chyby se kromě obsahu dokumentu mohou promítnout i do metadat:

4.4.1 Typografické chyby

Beall uvádí, že jednou z mnoha výhod digitálních objektů je, že v případě, kdy je chyba opravena v digitálním dokumentu, tak je narozdíl od tištěných verzí opravena navždy. Problém typografických chyb v online prostředí je často podceňován. Informační pracovník si nemusí uvědomit, že se dokument neobjevil ve vyhledávání kvůli překlepu, protože se předpokládá, že prohledávaná metadata reprezentují úplné a přesné informace. Mezitím v důsledku špinavých dat nebyly některé relevantní dokumenty ve výsledcích dotazu zahrnuty.

Existuje několik různých klasifikací typografických chyb. V základním rozsahu je lze rozlišit na chyby způsobené vynecháním, vložením, nahrazením a transpozicí znaku(Gardner, 1992)

Je však naděje, že jednotlivá slova, u kterých se typografické chyby v dokumentu vyskytnou, budou v dokumentu pravděpodobně někde napsány i správně. Vyhledávání frází nebo na základě blízkosti termínů tak nemusí být v některých případech nutně ovlivněno. Přesto je důležité tyto chyby vyhledávat a opravovat, zejména v metadatech digitálního objektu. Hlavně v případech, kdy metadata plní funkci zástupce samotného objektu. Pro některé typy objektů, jako jsou například obrázky nebo dokumenty bez dostupného plného textu, metadata plní funkci jediného přístupového bodu pro vyhledávací rozhraní a je tedy nutné zajistit jejich úplnou správnost (Beall, 2016)

4.4.2 Chyby způsobené skenováním a konverzí dat

Jedná se o poměrně nový typ chyby v digitálních objektech týkající se spíše kvality dat ve smyslu obsahu objektu. Mohou se ale objevit i v metadatech v případě, kdy jsou některá pole vyplňována automaticky na základě údajů ze skenů (např. čárové kódy nebo identifikátory z obálky zpracované pomocí optického rozpoznávání znaků). Tyto chyby jsou způsobeny chybným převodem textu z tištěného na digitální formát. Software použitý pro skenování může např. vložit mezery doprostřed slova, nebo špatně interpretovat písmeno "l" jako "i". Tyto chyby lze eliminovat podrobnou kontrolou výsledků skenování, což ale může být při velkých objemech dat prakticky nemožné. Tyto typy chyb se mohou také objevit v textových dokumentech při konverzích mezi formáty. V případě, kdy tyto chyby nejsou odstraněny, mohou ovlivnit indexaci a výsledky vyhledávání, protože snižují věrohodnost reprodukce kvality dat. Tato chyba se také objevuje častěji v případě jazyků s diakritikou nebo metadatových položek, ve kterých se mohou vyskytnout symboly.

4.4.3 Chyby způsobené funkcí Hledat & Nahradit

Tento typ chyby není až tak častý. Jedná se o případ, kdy jsou v případě automatického procesu chybně nahrazena slova nebo řetězce znaků za jiné. K tomu může dojít např. při použití softwaru pro korekci gramatiky nebo prediktivního zadávání textu. V rámci metadat k tomuto typu chyby většinou nedochází, pokud na ně tyto funkce nejsou přímo využity.

4.5 Proces ztotožňování v rámci vědeckých publikací a chybovost metadat

Pokud nelze publikace propojit pomocí jednoznačných identifikátorů je nutné použít ztotožňování záznamů. Proces ztotožňování v rámci vědeckých publikací je ve své podstatě velmi podobný jako například u ostatních typů publikací, objektů nebo osob. Jednotlivé záznamy reprezentují reálnou entitu, která je v záznamu popsána jednotlivými atributy. Dále je potřebné rozlišovat mezi ztotožňováním na úrovni celé publikace nebo pouze jejích částí. článku a na úrovni publikace.

V dnešní době většina článků podléhající recenznímu řízení a publikací vydané komerčními vydavateli mají přidělený identifikátor DOI, často i na úrovni kapitol. Naopak u šedé literatury, zejména konferenčních příspěvků, identifikátor často chybí. Zároveň také platí, že i v případech, kdy je DOI přiřazeno, nemusí být tato informace zachována ve všech verzích záznamů v různých systémech.

V případě, kdy nemáme k dispozici unikátní identifikátor, lze shodnost záznamů určit pomocí kombinace několika/řady atributů. Např. víme-li, že dvě publikace mají identický název, byly vydány ve stejném roce a na stejném místě, nicméně jejich autoři se jmenují Karel Novák a Karel Nivák, lze usoudit, že se jedná s vysokou pravděpodobností o překlep v příjmení a záznamy reprezentují stejnou knihu. Při strojovém zpracování můžeme tento postup napodobit použitím porovnávání znakových řetězců na základě odlišnosti vyjádřené metodami přibližné shody. Je-li např. editační vzdálenost pod určitou prahovou hodnotou, můžeme prohlásit, že záznamy reprezentují stejnou entitu.

4.6 Výběr údajů pro ztotožňování záznamů

Bez ohledu na to, zda pro ztotožňování záznamů použijeme pravděpodobnostní nebo vzdálenostní přístup, je potřeba vhodně vybrat atributy, na jejichž základě bude porovnávání probíhat. Z předchozích kapitol je zřejmé, že pokud chceme zvýšit kvalitu propojování záznamů, musíme použít vhodnou kombinaci polí, na jejichž základě budou záznamy porovnávány. Dále lze usuzovat, že počet těchto údajů také může hrát roli v přesnosti propojování, ale stejně tak může podstatně zpomalit tento proces a zvýšit komplexitu rozhodovacích pravidel.

Dále existuje mnoho způsobů jak dále usnadnit celý proces. V případě, kdy záznamy publikací obsahují jednoznačný identifikátor, lze množinu záznamů, jež vyžadují porovnání, podstatně zmenšit. Použitím autoritních dat v databázích můžeme podstatně zmenšit riziko záměny fyzických a právnických osob. Podobně je tomu i u ostatních polí za předpokladu, že jsou standardizovaná a platí pro ně vhodně zvolená omezení a požadavky na formátování.

V případě reálného použití přibližné shody znakových řetězců na metadatech vědeckých publikací se budou jednotlivá řešení měnit podle prostředí daného repozitáře či systému. Vzhledem k různým stupňům úplnosti záznamů je nutné vybrat relevantní sloupce, které jsou vhodné pro ztotožňování dokumentů. Tyto sloupce by měly splňovat požadavky, aby:

a) bylo podle nich možné buď identifikovat dokument, případně jej použít jako sekundární podklad pro rozhodování,
b) byly uvedené ve všech nebo skoro všech záznamech,
c) bylo je možné standardizovat bez ztráty významu.

Na základě těchto parametrů tedy připadají v úvahu tato nejběžnější pole:

Pole	**Výhody**	**Nevýhody**
Název publikace	Vysoká frekvence výskytu; variabilita řetězců; lze použít samostatně	Délka řetězce; malé změny mohou vést k velkým rozdílům
Autor publikace	Vysoká frekvence výskytu; variabilita řetězců; rychlost zpracování	Může být více údajů s proměnným pořadím v jednom poli; lze použít pouze v kombinaci s dalšími poli
Místo, rok vydání a nakladatel	Vysoká frekvence výskytu; rychlost zpracování	Lze použít pouze v kombinaci s dalšími poli; přestože se jedná o tři údaje, nemusí být v některých případech uvedeny jako separátní pole
Pomocné identifikátory	Nízká chybovost; unikátní; nemusí nutně vyžadovat přibližnou shodu; jasně identifikují objekt	Nemusí být vždy dostupné
Typ média, typ publikace	Dobré pro filtrování nerelevantních záznamů pro porovnávání, standardizované	Nemusí být vždy správně uvedeny, neidentifikují objekt jako takový
Údaje o periodicitě	Dobré pro filtrování nerelevantních záznamů pro porovnávání, standardizované	Neidentifikují objekt jako takový; nejsou dostupné u všech typů publikací
Technická metadata	Umožňují odlišit potenciální duplikáty	Nejsou vždy dostupná; neidentifikují objekt; velká variabilita jak v rámci obsahu, tak rozsahu obsažených údajů
Informace o konferencích	Umožňují odlišit potenciální duplikáty	Nejsou vždy dostupná; neidentifikují objekt; různé formy zápisu

Tabulka č. 1 Výhody a nevýhody běžných metadatových polí při ztotožňování záznamů

Pokud tedy budeme zvažovat použití metod přibližné shody znakových řetězců, budeme bilancovat mezi výpočetní náročností (primárně nás zajímají paměťové nároky a čas zpracování) a přesností. Čím více polí použijeme, tím roste doba, jak dlouho bude celý proces trvat. Na druhou stranu se snažíme minimalizovat počet chyb. Obecně vzato, ve velkých datasetech je snazší "rozpojit" chybně propojené záznamy, než dohledávat propojené záznamy, jež nebyly zachyceny.

5. Popis institucionálního systému ČVUT o aktuálním výzkumu

Pro správu dat o aktuálním výzkumu ČVUT používá vlastný systém nazvaný V3S. Ten slouží ke sdílení současných vědeckých informací, eviduje výsledky vědy a výzkumu (publikační výsledky, výsledky aplikovaného výzkumu) a další aktivity vědecko-výzkumných pracovníků ve vědecké komunitě. Tyto výsledky jsou následně odesílány do RIV (Rejstřík informací o výsledcích), případně jsou exportovány pro statistické analýzy, anebo jsou použity k interním hodnocením vědeckovýzkumné činnosti. Čtyři nejdůležitější moduly aplikace jsou: vyhledávání, vkládání a editace záznamů, statistika a import, export (viz následující podkapitoly). (V3S, 2019)

5.1 Vyhledávání

Aplikace umožňuje základní i rozšířené vyhledávání informací o vložených záznamech. Vyhledávací dotaz lze upřesnit množstvím nabízených parametrů, jejichž počet není limitován, a řetězce lze nahrazovat zástupnými znaky. Uživatelé mají možnost ukládat a exportovat svá vyhledávání a také filtry v rámci xml souborů. Samotné výsledky lze vzestupně i sestupně řadit na základě použitých parametrů vyhledávání. Některé parametry lze automaticky vyplnit údaji přihlášeného uživatele pomocí ikonek “moje osoba” a “moje pracoviště”.

Výchozí filtrovací kritéria pro vyhledávání jsou:

- **Název, anotace:** pro vyhledání výsledku/ů je možné zadat část názvu nebo anotace hledaného/ných výsledků
- **Součást:** výběr součásti ČVUT z rozbalovací nabídky
- **Pracoviště:** výběr z rozbalovací nabídky závislé na výběru součásti, nebo přes ikonu „moje pracoviště“; zadáním ID nebo názvu přes našeptávač
- **Autor:** výběr možný přes našeptávač, ikonu „moje osoba“ či vyhledávání osob
- **Kontrola RIV:** výběr z rozbalovací nabídky se seznamem všech možných stavů kontroly RIV
- **ID záznamu:** vyhledání výsledku dle zadání jeho ID (přidělené systémem)
- **Stav záznamu:** výběr z rozbalovací nabídky se seznamem všech možných stavů záznamů výsledku
- **ISBN:** textový řádek, výběr výsledku proběhne dle zadaného ISBN

- **ISSN:** textový řádek, výběr výsledku proběhne dle zadaného ISSN
- **Od roku, Do roku:** zadaný údaj se porovnává s polem Rok vydání v záložce Popisu výsledku

Systém je také vybaven tlačítkem "+", které umožňuje přidávat další pole stejného typu a vyhledávat podle více hodnot současně. Další nabízené filtry umožňují zúžení vyhledávání na informace ohledně vydání (forma, vydavatel, místo / stát), konference (pořadatel, místo konání, datum konání), na informace externích databází (identifikátory: WoS a Scopus, DOI; systémové číslo katalogu, citace, identifikátor citující publikace), údaje ohledně vložení a úprav záznamů, případně jejich licence, hodnocení a rozpory.

Kromě samotných záznamů lze také vyhledávat citace následujícími filtry:

- **Autor výsledku:** výběr možný přes našeptávače, ikonu "moje osoba" či vyhledávání osob
- **Součást:** výběr části ČVUT z rozbalovací nabídky
- **Pracoviště:** výběr z rozbalovací nabídky závislé na výběru součásti, ikonu "moje pracoviště, zadáním ID nebo názvu přes našeptávač.
- **Od roku, Do roku:** zadaný údaj porovnává s polem Rok vydání v záložce Popisu výsledku
- **Typ výsledku:** výběr z rozbalovací nabídky se seznamem všech možných typů citovaného výsledku
- **Citováno od roku, Do roku:** zadaný údaj se porovnává s rokem citace
- **Zahrnout autocitace:** filtr citací, kdy autor cituje sám sebe a své práce
- **Významy citací:** skupina zaškrtávacích políček, vyjadřuje druh a původ citace. Umožňuje výběr více nebo i všech hodnot.

Dále lze vyhledávat také prodané licence, organizace a uznání vědeckou komunitou.

5.2 Vložení, editace a prohlížení záznamů

Kromě vyhledávání záznamů aplikace plní i funkci editoru. Záznamy se stahují automaticky z WoS a SCOPUS, případně se vkládají přes vlastní položku v menu, kde uživatel vybere z nabízených možností typ dokumentu a vyplní informace. Zadávací formulář výsledku disponuje povinnými a nepovinnými poli. Povinné položky jsou označeny červenou hvězdičkou, položky povinné pro uložení výsledku do stavu dokončeno jsou označeny modrou hvězdičkou. Jednotlivá pole jsou vybavena nápovědou s informacemi a pokyny pro jejich vyplnění. Rozpracované záznamy je možné také uložit bez nutnosti odeslání. Systém je také schopen rozpoznat situaci, kdy uživatel opouští stránku bez uložení své práce a zobrazí dialogové okno s dotazem, aby se předešlo ztrátě neuložené práce. Po vytvoření záznamu je přiděleno identifikační číslo, které je unikátní v rámci aplikace.

5.3 Statistiky a hodnocení

Tato komponenta umožňuje několik různých výstupů: souhrnná hodnocení ČVUT na základě publikací a citací na Web of Science, podle H-Indexu, hodnocení RVVI (Rada pro výzkum, vývoj a inovace), podklady pro habilitace a jmenování profesorem, body za impaktované publikace a citace a fakultní hodnocení.

5.4 Import, Export

Tento modul v prvé řadě umožňuje řešitelům akcí (projekt či smluvní výzkum) a vedoucím nahrávat do systému jednotlivé dokumenty. Zároveň je možné generovat širokou škálu reportů jako např.: interní evaluace pro oddělení některých fakult, hlášení sledující plnění povinností akademických pracovníků nebo export pro databázi RIV. (Dvořák, 2019)

Systém V3S primárně čerpá externí data o publikacích ze dvou zdrojů, a to z citačních databází Web of Science a Scopus.

6. Vyhodnocení úspěšnosti různých metod ztotožňování

Primárním cílem výzkumu bylo potvrdit vhodnost metod založených na vzdálenosti znakových řetězců, navíc byla práce poměrně limitována hardware nároky. Název publikace se hned po jednoznačných identifikátorech jeví jako nejspolehlivější způsob, jak na minimálním počtu použitých polí porovnat a propojit záznamy[2]. Aby bylo možné výzkum provést v rozumném časovém období, bylo porovnávání rozšířeno o blokaci na základě roku vydání. V reálném nasazení na množině bez dostupných identifikátorů by bylo vhodné porovnávací parametry rozšířit o autory publikace a nakladatelské údaje. Ostatní údaje je poté možno použít pro kontrolu u záznamů, kde není dostatečná míra důvěry pro propojení záznamů, ale zároveň nelze vyloučit jejich možné propojení.

6.1 Cíle výzkumu

Pro praktické otestování použití metody přibližné shody v rámci metadat vědeckých publikací byly stanoveny následující cíle práce:

1) Vytvoření nástroje pro ověření vhodnosti aplikace vybraných metrik na poskytnutém datasetu a interpretaci výsledků formou precision-recall grafů či jiných dostupných forem evaluace propojovací kvality a komplexity;

2) Ověřit chování vytvořeného nástroje na validačním datasetu bez dostupných DOI identifikátorů formou ruční kontroly.

6.2 Metodika výzkumu

Účelem výzkumu je porovnat použití metrik popsaných v kapitole 3 v procesu ztotožňování záznamů na základě jejich metadat, tj.aplikace/použití Levenštejnovy vzdálenosti, Jarovy vzdálenosti, Jaro-Winklerovy vzdálenosti, kosinové vzdálenosti q-gramů a Jaccardova koeficientu. Za účelem získání co největší množiny vzdáleností bylo porovnávání provedeno na názvech jednotlivých publikací. Jako metoda porovnávání záznamů pro daný dataset byla zvolena binární klasifikace párů založená na prahové hodnotě.

[2] DOI identifikátory jsou poté použity pro ověření správnosti propojení na trénovací množině

Pro Jaro-Winklerovu vzdálenost byl stanoven standardní penalizační faktor p=0.1 (Winkler, 1990). Kosinová vzdálenost q-gramů a jaccardův koeficient byly měřeny pro q-gramy o velikosti 3 a 4.

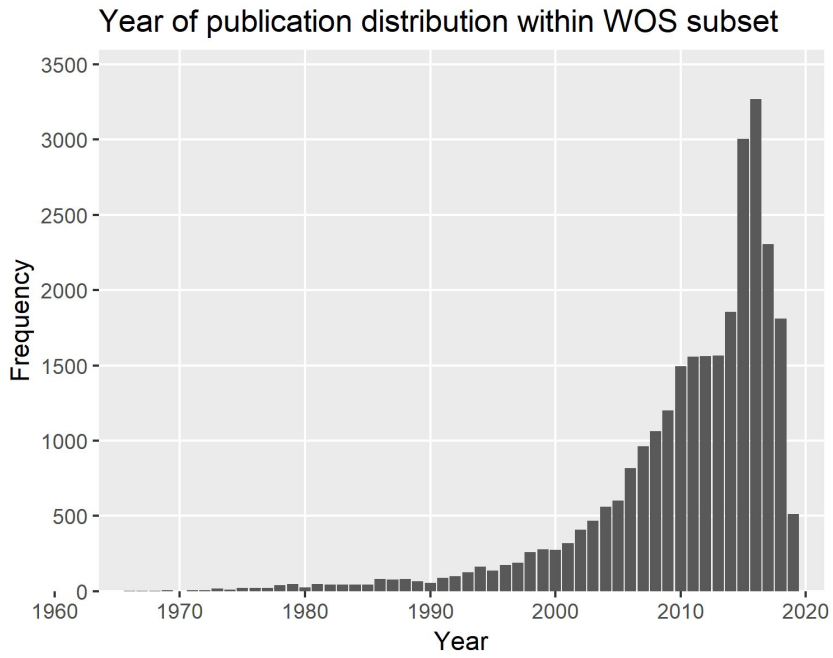
6.3. Data

Zdrojová datová sada pochází ze systému V3S spravovaného Českým vysokým učením technickým v Praze (viz kap. 5). Samotný dataset má standardní formát tabulky, přičemž každý řádek reprezentuje jeden metadatový záznam publikace. Dataset obsahoval 27 935 záznamů z Web of Science a 29 881 z databáze Scopus. Obr. č. 9 a 10 uvádí počty záznamů podle roku vydání.

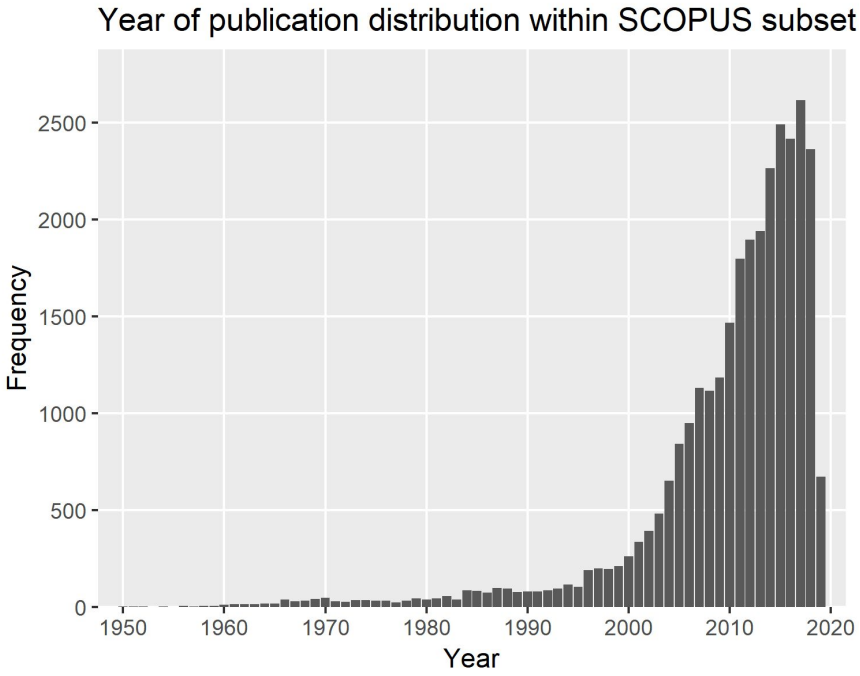
V těchto množinách 15 381 záznamů z WoS a 22 039 ze Scopus obsahovalo DOI. Shodné DOI mezi záznamy z WoS a Scopus propojilo 12 467 párů, které jsou použity jako trénovací množina. V případě 127 z těchto párů se lišil rok publikace. 12 340 (99 %) párů mělo stejný rok publikace, u všech ostatních párů byla odchylka nižší než +-3 roky. Z párů propojených podle DOI se pouze u 6 632 (53 %) shodoval název publikace.

Licenční podmínky poskytovatelů databází Web of Science a Scopus bohužel neumožňují zveřejnění zdrojové datové sady. Ta je archivována na Výpočetním a informačním centru ČVUT v Praze.

Zpracovaná data vč. popisu struktury jsou zpřístupněna na:
https://doi.org/10.5281/zenodo.3785363.



Obr. č. 8: Distribuce publikací z Web of Science podle roku vydání.



Obr. č. 9: Distribuce publikací ze Scopus podle roku vydání.

6.4. Nástroje

Pro zpracování byl zvolen programovací jazyk, prostředí určené pro statistickou analýzu dat a jejich grafické zobrazení **R.**[3]

Plnou dokumentaci a zdrojový kód programů vytvořených ke zpracování údajů pro tuto diplomovou práci lze nalézt na adrese https://github.com/jdobiasovsky/metric-test.

Celý proces byl rozdělen na tři hlavní části: příprava, zpracování a vizualizace.

6.5 Příprava datasetu

Dataset obsahoval záznamy s datem publikace od roku 1950 do 2019. Nejvíce záznamů pochází z období za posledních 10 let, což rozhodovalo při vytyčení finálních skupin, na kterých se bude efektivita metrik porovnávat. Aby byla využita data v co největším rozsahu, byly vybrány 3 skupiny: celý rozsah datasetu (1950 - 2019), posledních 10 let (2009 - 2018) a poslední 3 roky (2016 - 2018).

Záznamy jsou také rozděleny podle zdrojové databáze na dva separátní datasety obsahující standardizované záznamy z WoS a Scopus. Příklad výsledného záznamu vypadá následovně[4]:

$ PUBLICATION <chr> "43178"
$ YEAR <dbl> 2000
$ DOI_CODE <chr> "10.1002/1099-0682"
$ TITLE <chr> “Cognitive research in information science: implications for design”
$ ABSTRACT <chr> NA
$ SOURCE <chr> “Information science journal”
$ PUBLISHER <chr> "wiley"
$ PUBLISHER_LOCATION <chr> NA
$ CONFERENCE_NAME <chr> “Annual ILS conference 2000”
$ AUTHORS <chr> “Němcová B.”, “Vacková K.”, “Kára J.”

[3] https://www.r-project.org/

[4] jednotlivé sloupce záznamu jsou prezentovány jako řádky pro lepší přehled

$ AUTHORS_CTU <chr> "Kára J."

Záznam je pro lepší strojové zpracování poté standardizován - zápis je převeden na pouze malá písmena, jsou odstraněny mezery a speciální znaky. Záznam poté vypadá následovně:

$ PUBLICATION <chr> "43178"
$ YEAR <dbl> 2000
$ DOI_CODE <chr> "10.1002/1099-0682"
$ TITLE <chr> "cognitiveresearchininformationscienceimplicationsfordesign"
$ ABSTRACT <chr> NA
$ SOURCE <chr> "informationsciencejournal"
$ PUBLISHER <chr> "wiley"
$ PUBLISHER_LOCATION <chr> NA
$ CONFERENCE_NAME <chr> "annualilsconference2000"
$ AUTHORS <chr> "němcovab", "vackovak", "karaj"
$ AUTHORS_CTU <chr> "karaj"

Na základě poznatků z kapitoly 4.6 pak probíhá porovnávání názvů publikací u záznamů, jež odpovídají filtračním kritériím podle roku publikování.

6.6 Zpracování

Zpracování probíhá porovnáváním záznamů ze Scopus se záznamy z databáze WoS (Web of Science). Vzhledem k hardwarovým a časovým limitům, nebylo možné dataset jako takový zpracovat metodou porovnání každého záznamu ze Scopus s každým ze záznamů z WoS.
Záznamy jsou ve fázi indexace blokovány formou upraveného invertovaného indexu vytvořeného na základě jejich roku vydání. Tvorba indexu a následovných párů vypadá na příkladu následovně:

1) Program zaznamená aktuálně zpracovávaný rok 2000.
2) Z databáze Scopus jsou nejprve vybrány cílové záznamy pro tento rok.

3) Z databáze WoS jsou vybrány kandidátní záznamy v tříletém rozpětí. Tedy v tomto případě by se jednalo o záznamy z roku 1999-2001.

4) Výsledkem jsou dvě množiny záznamů.

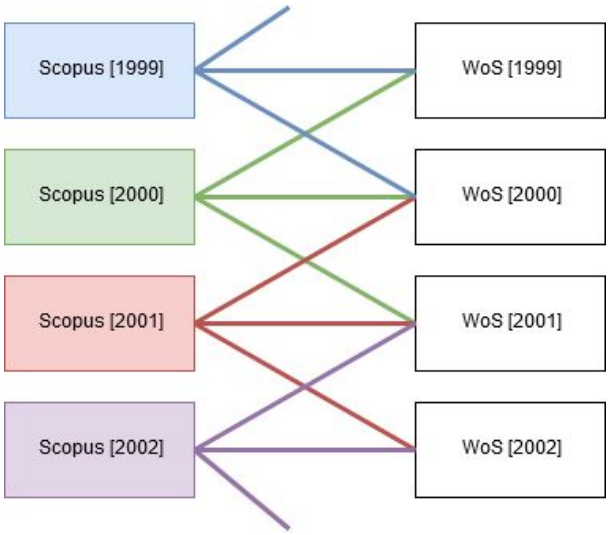
db_WOS [2000]
záznam 1
záznam 2

db_SCOPUS [1999-2001]
záznam a
záznam b
záznam c

5) Je vytvořena tabulka porovnávacích párů, kde každý záznam z databáze Scopus je porovnán se všemi záznamy z databáze WoS v daném bloku. Jednotlivé řádky zde reprezentují porovnávací pár.

db_WOS [2000]	**db_SCOPUS [1999-2001]**
záznam 1	záznam a
záznam 1	záznam b
záznam 1	záznam c
záznam 2	záznam a
záznam 2	záznam b
záznam 2	záznam c

6) Program ukládá tabulku do úložiště a přechází do další iterace. Tedy záznamy pro rok 2001 z databáze Scopus a záznamy z let 2000-2002 pro WoS. Tím, že jsou použity vždy záznamy pouze z jednoho roku pro Scopus, je zajištěno, aby se neprováděly stejné porovnávací operace ve více blocích, a překryvy u databáze WoS zvyšují pravděpodobnost propojení.



Obr. č. 10: Demonstrace překryvu skupin porovnávaných záznamů

Následuje výpočet vzdáleností textových řetězců. Pro šetření výpočetního výkonu jsou nejprve vygenerovány páry a až poté pomocí knihovny stringdist[5] vypočítány hodnoty vzdálenosti pro jednotlivé sloupce. Výsledkem zpracování je tabulka obsahující DOI, rok vydání, id každého záznamu z páru a hodnoty vzdálenosti pro jednotlivé sloupce, které byly zahrnuty v porovnávání. Hodnoty jsou standardizovány tak, aby reprezentovaly vzdálenost v intervalu <0, 1>, kde 0 jsou identické řetězce a 1 jsou naprosto odlišné řetězce.

```
> open_data("./data/precision_recall_filtered_06/lv_2018.csv")
# A tibble: 14,115 x 8
      X1   ID1 YEAR1 DOI1                      ID2 YEAR2 DOI2                      TITLE
   <dbl> <dbl> <dbl> <chr>                   <dbl> <dbl> <chr>                     <dbl>
 1     1 50773  2018 10.1002/acm2.12238      52368  2018 10.1002/acm2.12238      0
 2  4630 54240  2018 10.1002/acn3.618        54563  2018 10.1002/acn3.618        0
 3  9260 54596  2018 10.1002/adem.201800182 54700  2018 10.1002/adem.201800182 0
 4 13894 53131  2018 10.1002/asmb.2337       53877  2018 10.1002/asmb.2337       0
 5 18524 55929  2018 10.1002/asna.201813363 56070  2018 10.1002/asna.201813363 0
 6 23153 54325  2018 10.1002/asna.201813498 54445  2018 10.1002/asna.201813498 0.837
 7 27782 54328  2018 10.1002/asna.201813507 54450  2018 10.1002/asna.201813507 0
 8 27784 54328  2018 10.1002/asna.201813507 54447  2018 10.1002/asna.201813510 0.278
 9 27786 54328  2018 10.1002/asna.201813507 54446  2018 10.1002/asna.201813513 0.417
10 32411 54330  2018 10.1002/asna.201813508 54453  2018 10.1002/asna.201813508 0
# ... with 14,105 more rows
```

Obr. č. 11: Příklad výsledkové tabulky s údaji o porovnávaných záznamech a výsledné míře shody znakových řetězců v názvu.

Výstupem zpracování procesu jsou soubory obsahující výsledky pro jednotlivé skupiny a metriky, které bylo možno kombinovat podle potřeb zobrazení.

6.7 Vizualizace výsledků

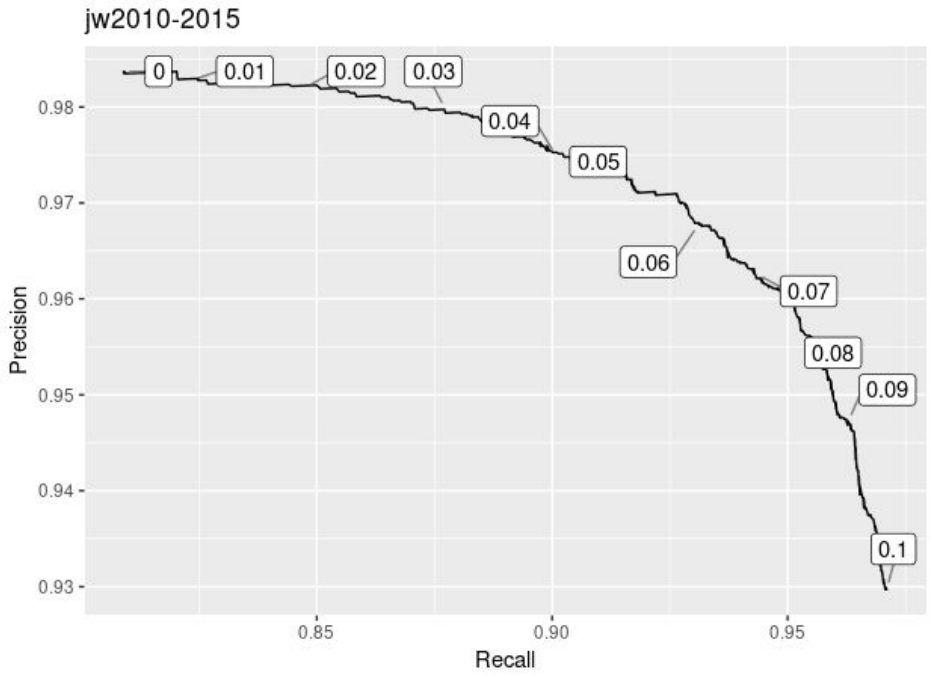
Pro vizualizaci syrových výsledků byla vytvořena funkce umožňující na základě uživatelem definovaného prahu a hodnot DOI klasifikovat propojení jednotlivých párů (true positive, true negative, false positive, false negative). Na základě počtů jednotlivých klasifikací lze poté určit hodnoty preciznosti, výtěžnosti a F-míry. Výsledky hodnot preciznosti, výtěžnosti a F-míry byly vypočítány pro prahy v rozsahu <0, 0.6 >, s krokem po 10^{-4}. Za použití knihovny ggplot[6] bylo možné následně vykreslovat grafy zobrazující vztahy mezi jednotlivými měřenými hodnotami.

[5] https://cran.r-project.org/web/packages/stringdist/index.html

[6] https://ggplot2.tidyverse.org/

```
> generate_results_exploratory(open_data(input = "./data/precision_recall_filtered/lv_2005.csv"), "TITLE")
   Treshold Precision    Recall  Fmeasure  TP FP FN
1      0.10 0.9729730 0.9191489 0.9452954 216  6 19
2      0.09 0.9729730 0.9191489 0.9452954 216  6 19
3      0.08 0.9726027 0.9063830 0.9383260 213  6 22
4      0.07 0.9720930 0.8893617 0.9288889 209  6 26
5      0.06 0.9719626 0.8851064 0.9265033 208  6 27
6      0.05 0.9714286 0.8680851 0.9168539 204  6 31
7      0.04 0.9711538 0.8595745 0.9119639 202  6 33
8      0.03 0.9708738 0.8510638 0.9070295 200  6 35
9      0.02 0.9705882 0.8425532 0.9020501 198  6 37
10     0.01 0.9696970 0.8170213 0.8868360 192  6 43
11     0.00 0.9695431 0.8127660 0.8842593 191  6 44
```

Obr. č. 12: Výsledková tabulka pro prahové hodnoty. (Pro výzkum byla použita vyšší granularita.)



Obr. č. 13: Demonstrace vykreslení hodnot prahu na osách pro Preciznost a Míru výtěžnosti.

6.8 Ruční kontrola na validačním vzorku bez dostupných identifikátorů

Na základě analýzy dosažených hodnot propojovací kvality a komplexity na záznamech z posledních tří let je zvolena optimální prahová hodnota pro jednotlivé algoritmy, u které bylo dosaženo nejvyšší F-míry.

Pro ruční kontrolu je vybrán náhodný vzorek 100 propojených párů záznamů pro každou z vybraných metrik. V daném propojeném páru buď oba, nebo alespoň jeden záznam neobsahují identifikátor DOI. Autor poté na základě dostupných informací z ostatních polí manuálně ověří, zda-li se jedná o pravdivě pozitivní výsledek (true positive) nebo o výsledek falešně pozitivní (false positive). Předpokládá se, že poměr pravdivých a nepravdivých propojení propojení by měly v tomto validačním vzorku být přibližně stejné jako u trénovací množiny.

7. Výsledek hodnocení úspěšnosti různých metod ztotožňování

Pro sjednocení programového výstupu jsou zavedeny následující zkratky:

lv - Levenshteinova vzdálenost

jaro - Jarova vzdálenost

jw - Jaro-Winklerova vzdálenost (penalizační faktor p=0.1)

cosine3 - kosinová vzdálenost pro q-gramy o velikosti 3 (trigramy)

cosine4 - kosinová vzdálenost pro q-gramy o velikosti 4 (quadrigramy)

jaccard3 - Jaccardův koeficient pro q-gramy o velikosti 3

jaccard4 - Jaccardův koeficient pro q-gramy o velikosti 4

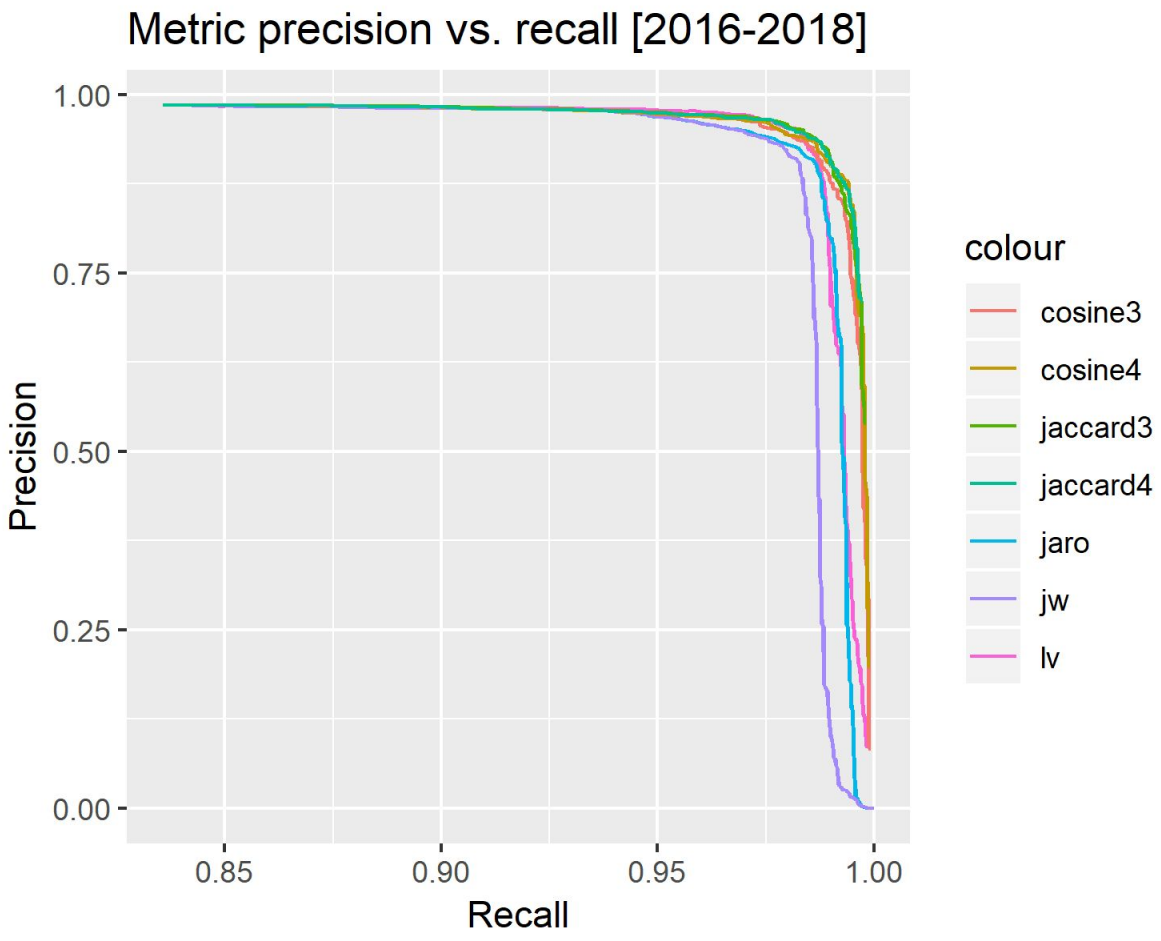
7.1 Vizualizace porovnání propojovací kvality jednotlivých metrik

V následující sekci jsou pro každou skupinu vykresleny 2 grafy.

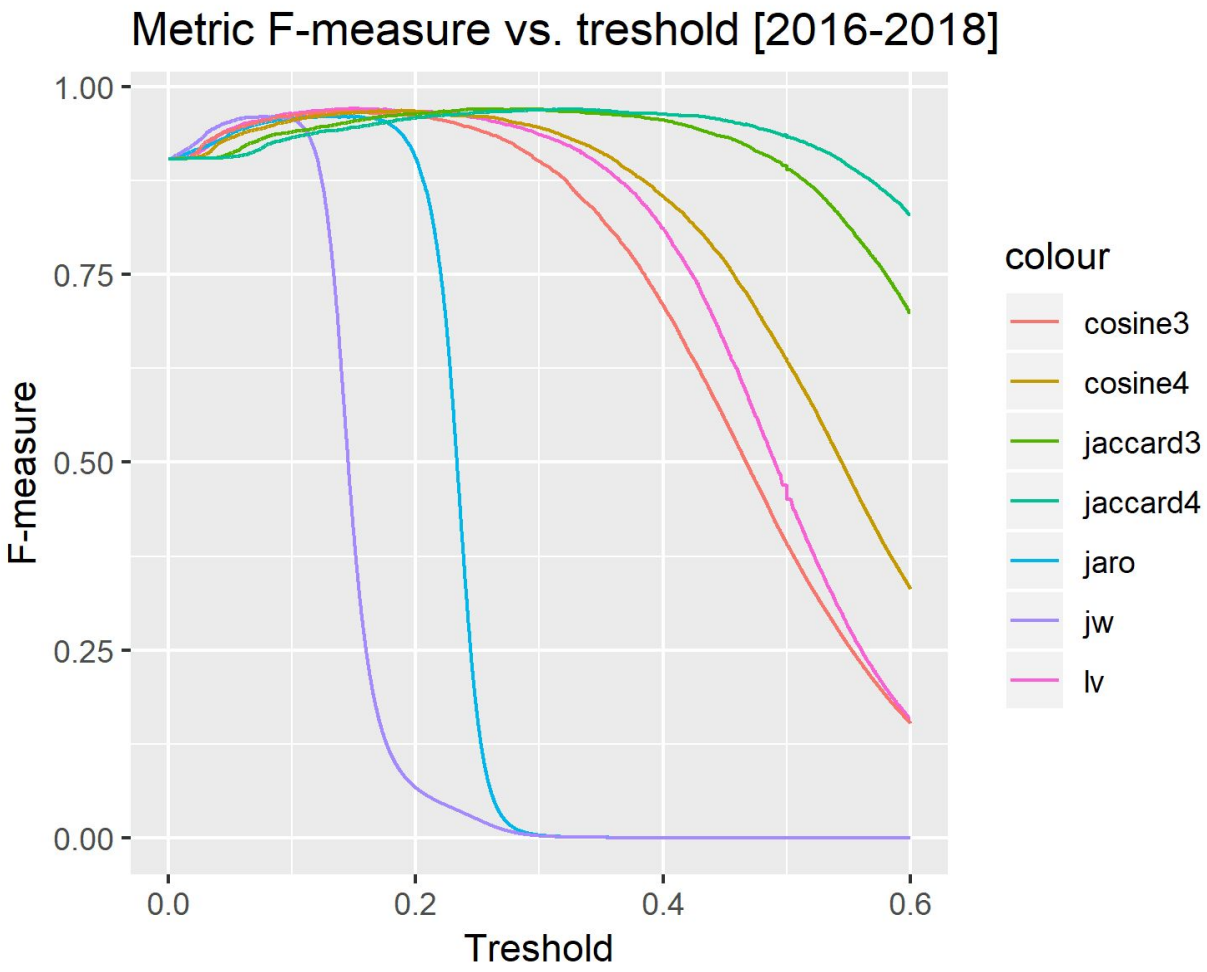
První graf porovnává preciznost a výtěžnost v daném období. Práh, u kterého sledujeme vysokou výtěžnost ale nízkou preciznost, vrací mnoho propojení, ale většina z nich je špatná. Naopak u prahu s vysokou precizností ale nízkou výtěžností vrací velmi málo výsledků, ale zato je většina z nich správně propojená. V ideálním případě tedy usilujeme o takový práh, jenž bude vracet vysoké hodnoty výtěžnosti i preciznosti.

Druhý graf vyobrazuje F-míru pro danou hodnotu prahu. Můžeme tedy sledovat, pro kterou hodnotu prahu začne postupně klesat preciznost metriky a tím pádem i její efektivita.

7.1.1 Skupina pro roky 2016-2018

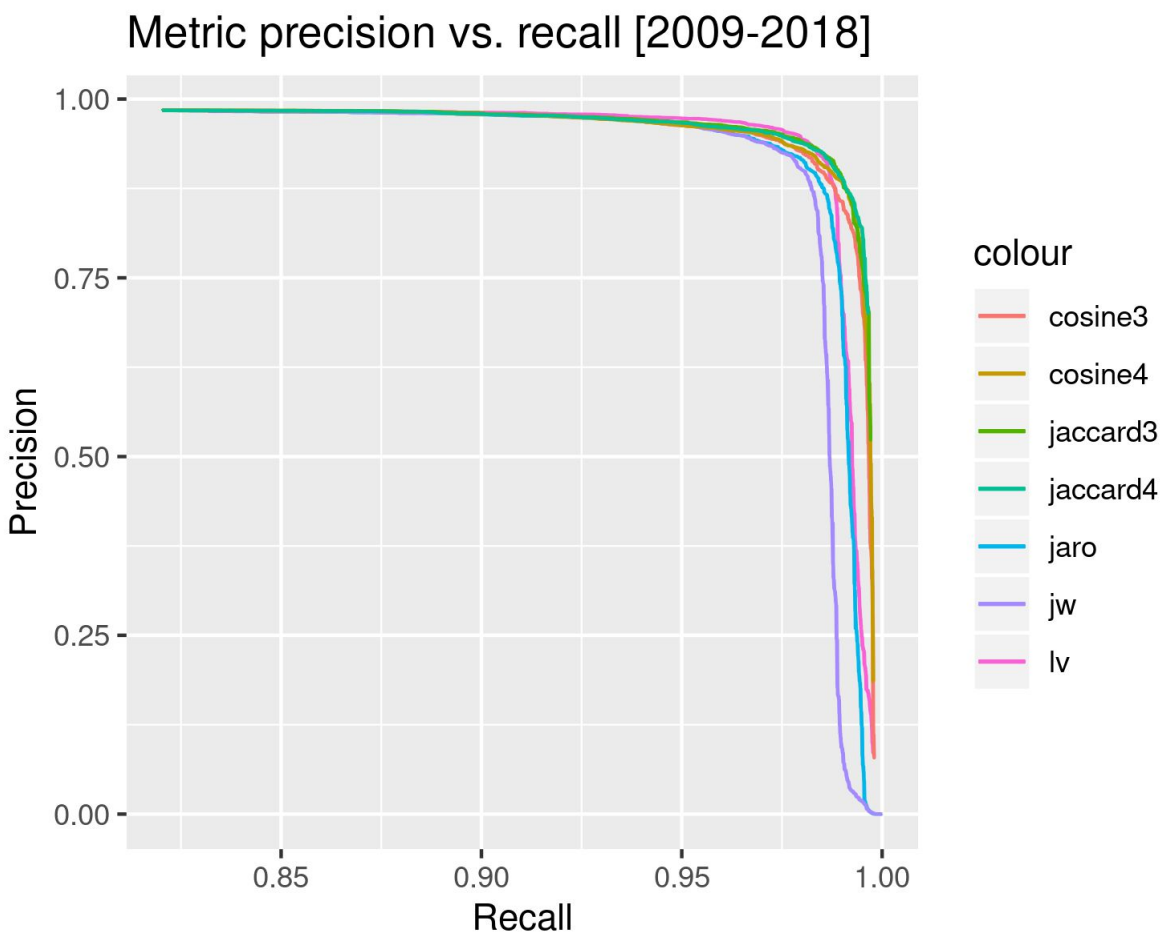


Obr. č. 14: Graf porovnávající hodnoty preciznosti a výtěžnosti jednotlivých metrik na publikacích s datem vydání mezi roky 2016-2018.

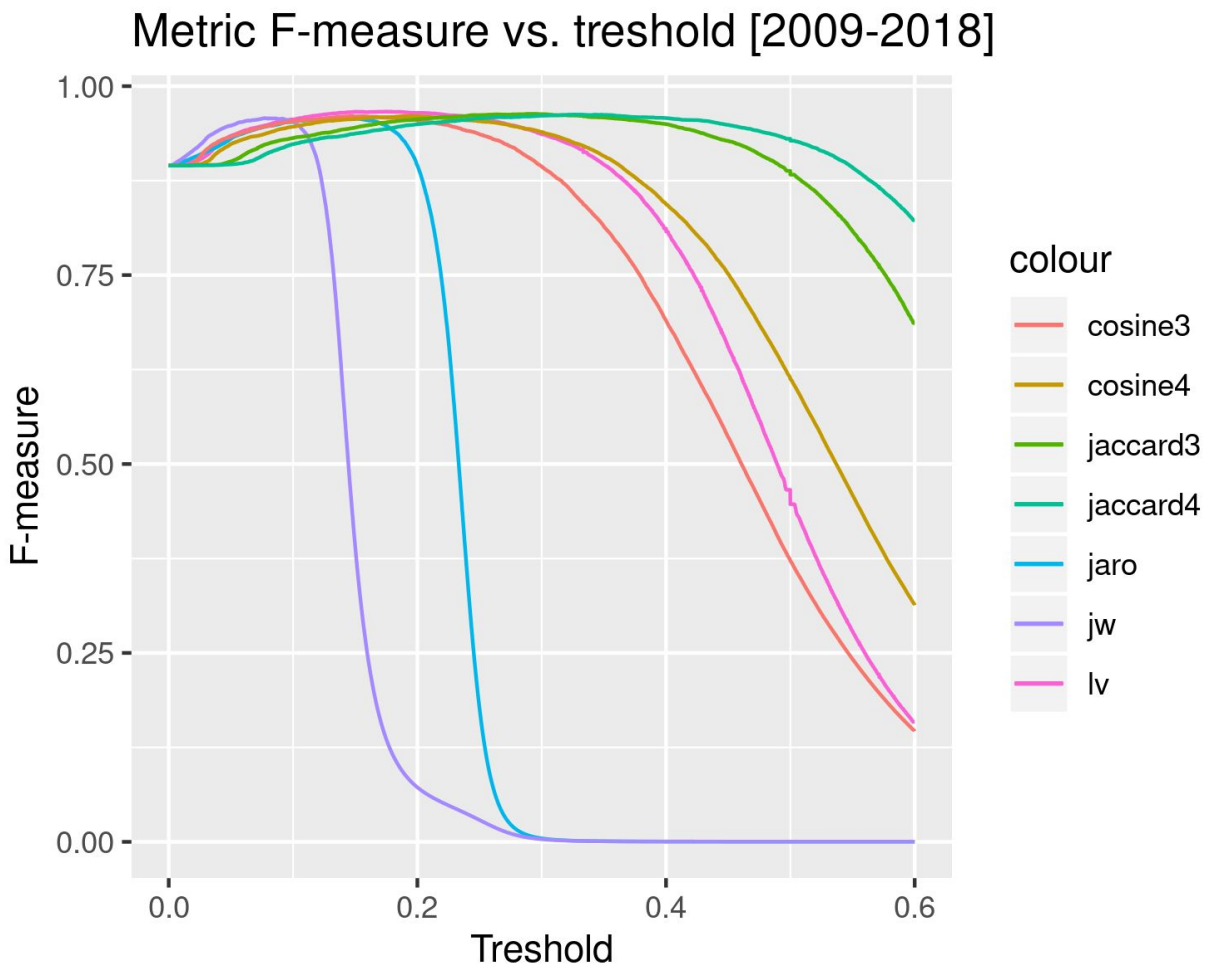


Obr. č. 15: Graf porovnávající hodnoty F-míry jednotlivých metrik na daných prazích u publikací s datem vydání mezi roky 2016-2018.

7.1.2 Skupina pro roky 2009-2018

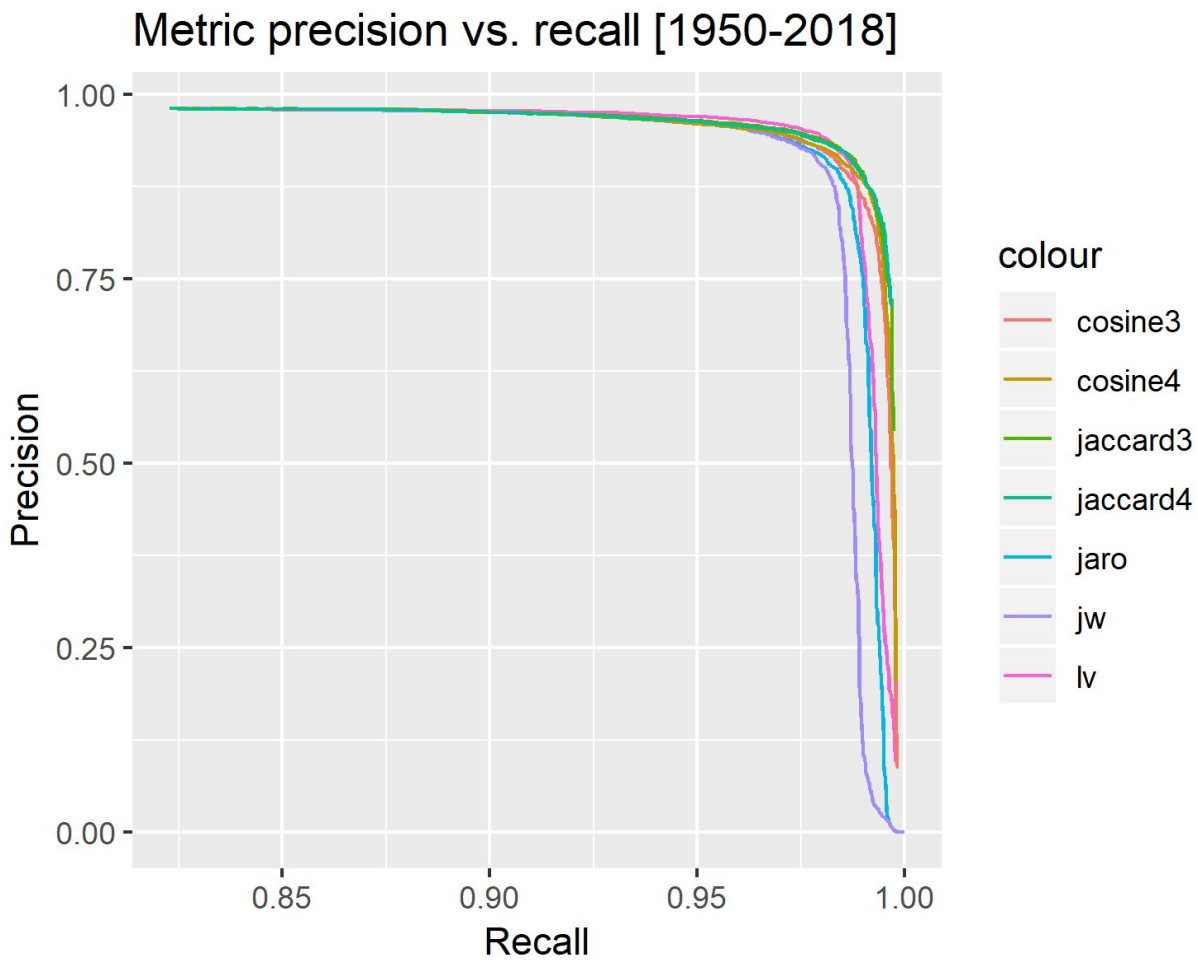


Obr. č. 16: Graf porovnávající hodnoty preciznosti a výtěžnosti jednotlivých metrik na publikacích s datem vydání mezi roky 2009-2018.

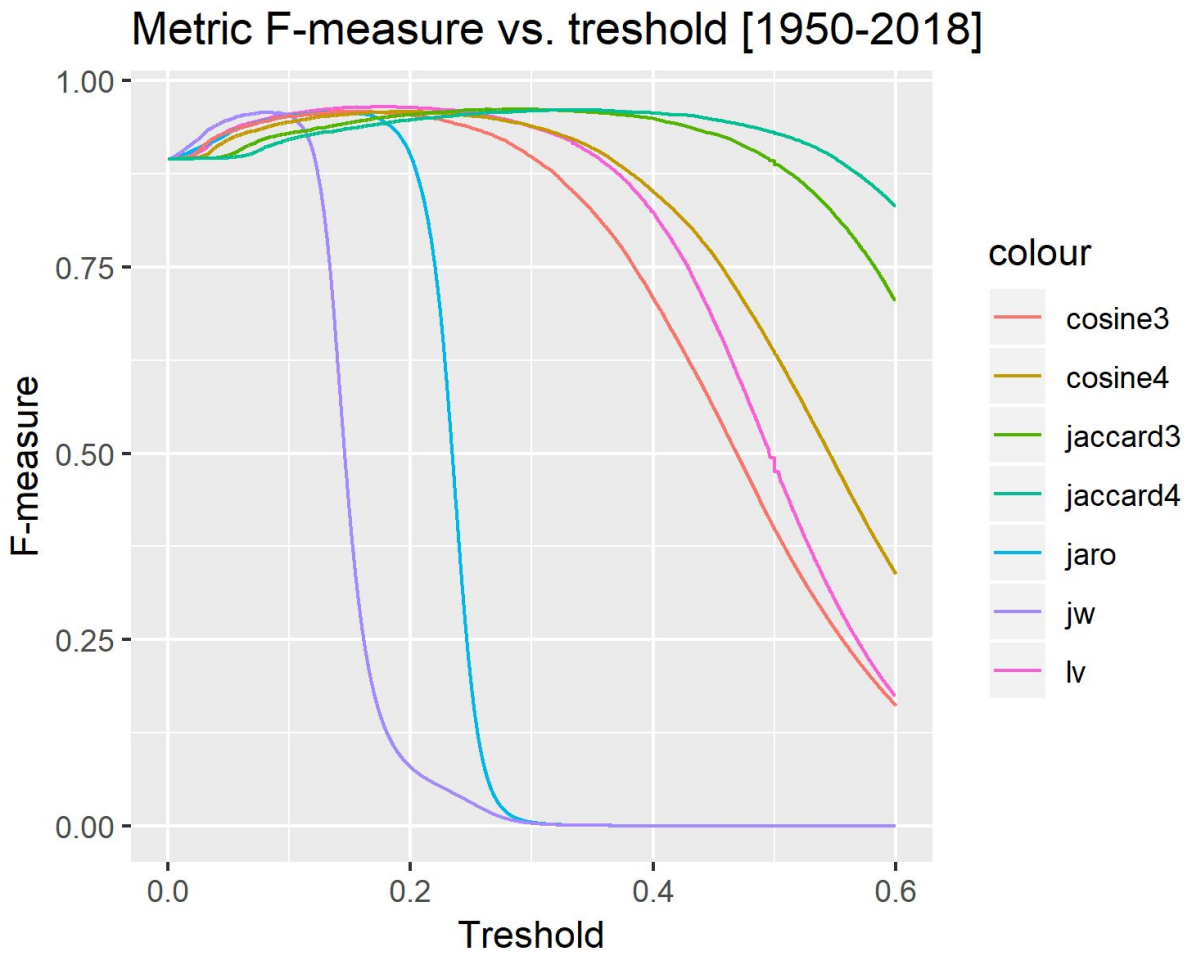


Obr. č. 17 Graf porovnávající hodnoty F-míry jednotlivých metrik na daných prazích u publikací s datem vydání mezi roky 2009-2018.

7.1.3 Skupina pro roky 1950-2018



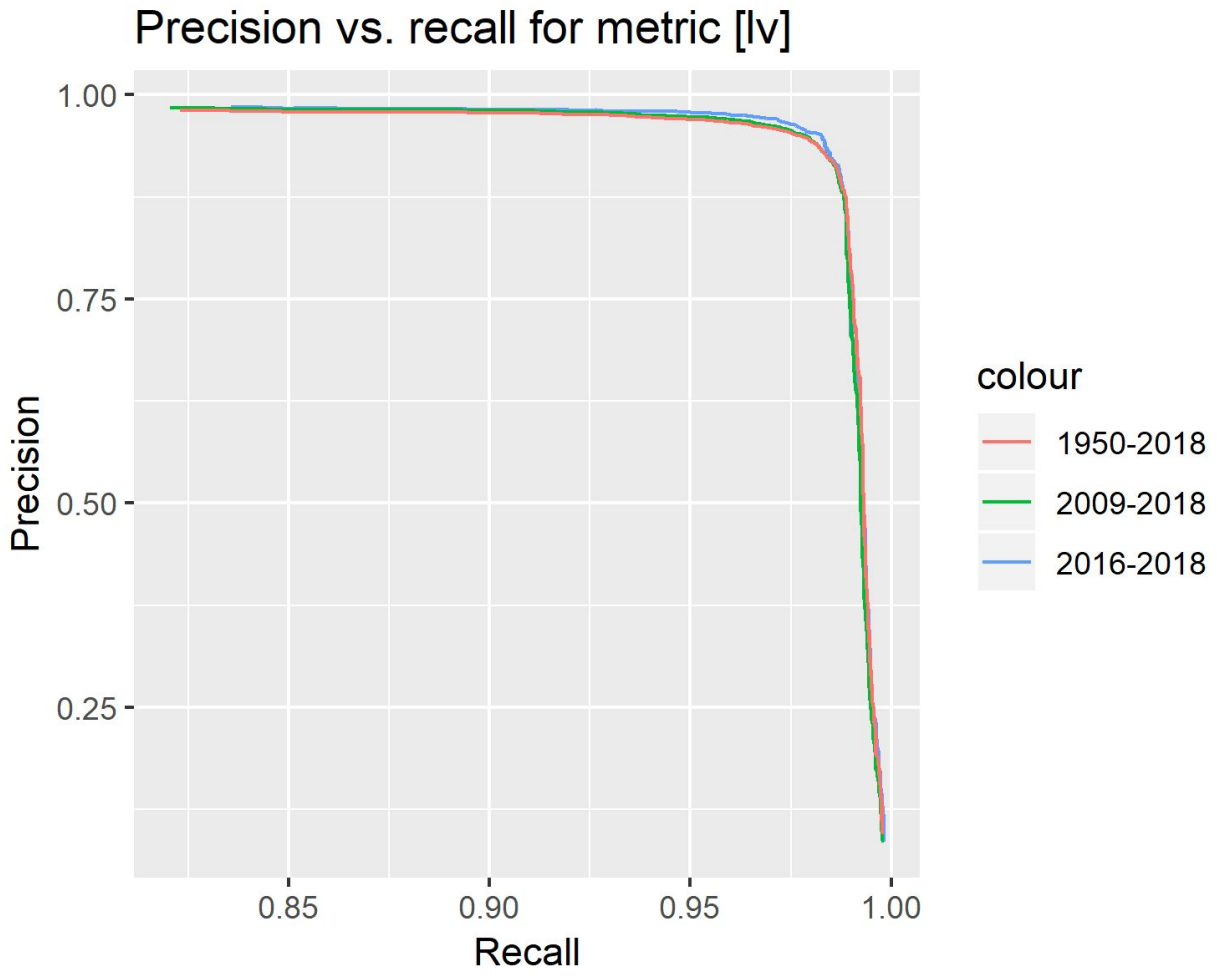
Obr. č. 18: Graf porovnávající hodnoty preciznosti a výtěžnosti jednotlivých metrik na publikacích s datem vydání mezi roky 1950-2018.



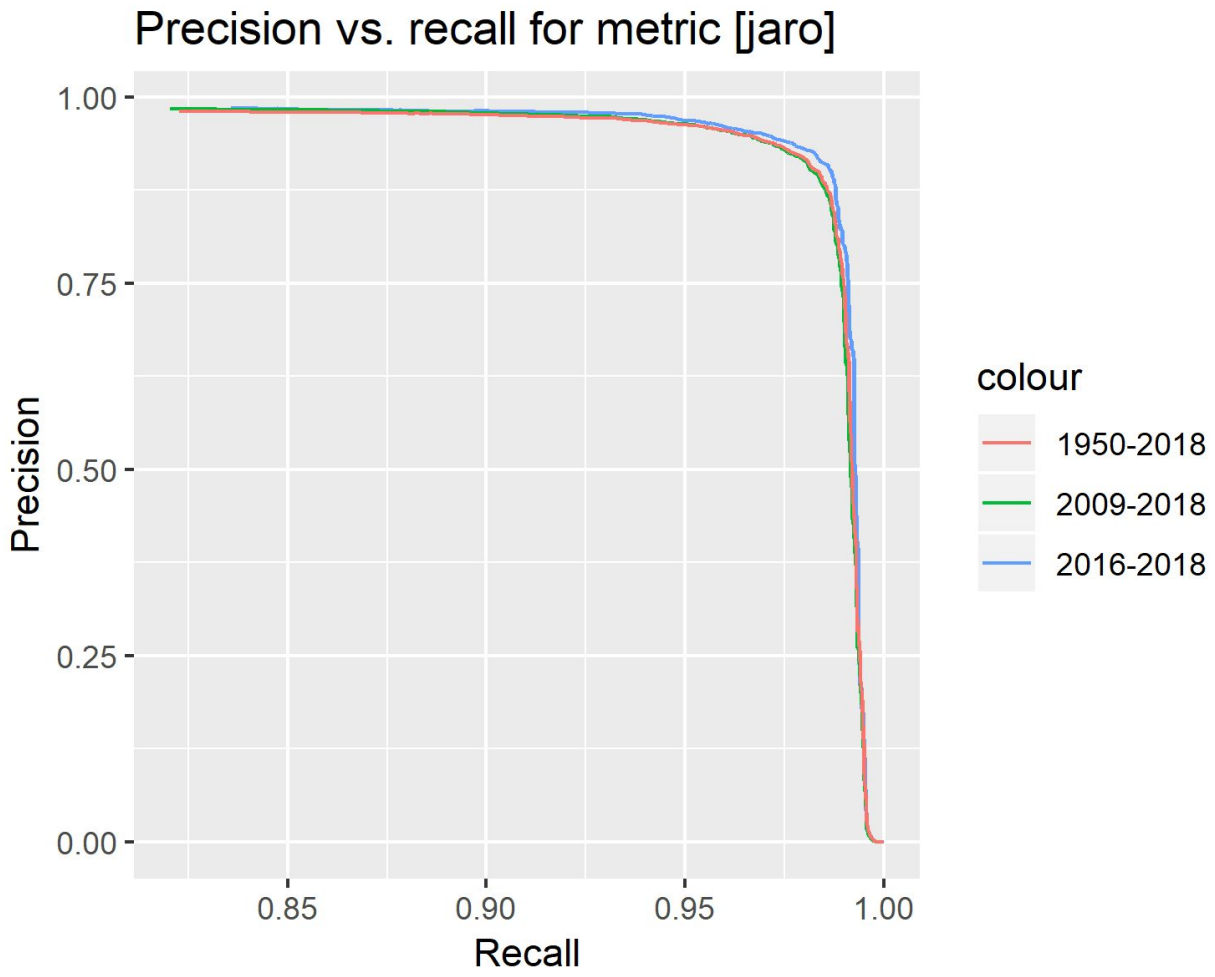
Obr. č. 19 Graf porovnávající hodnoty F-míry jednotlivých metrik na daných prazích u publikací s datem vydání mezi roky 1950-2018.

7.2. Kontrola robustnosti řešení mezi jednotlivými obdobími

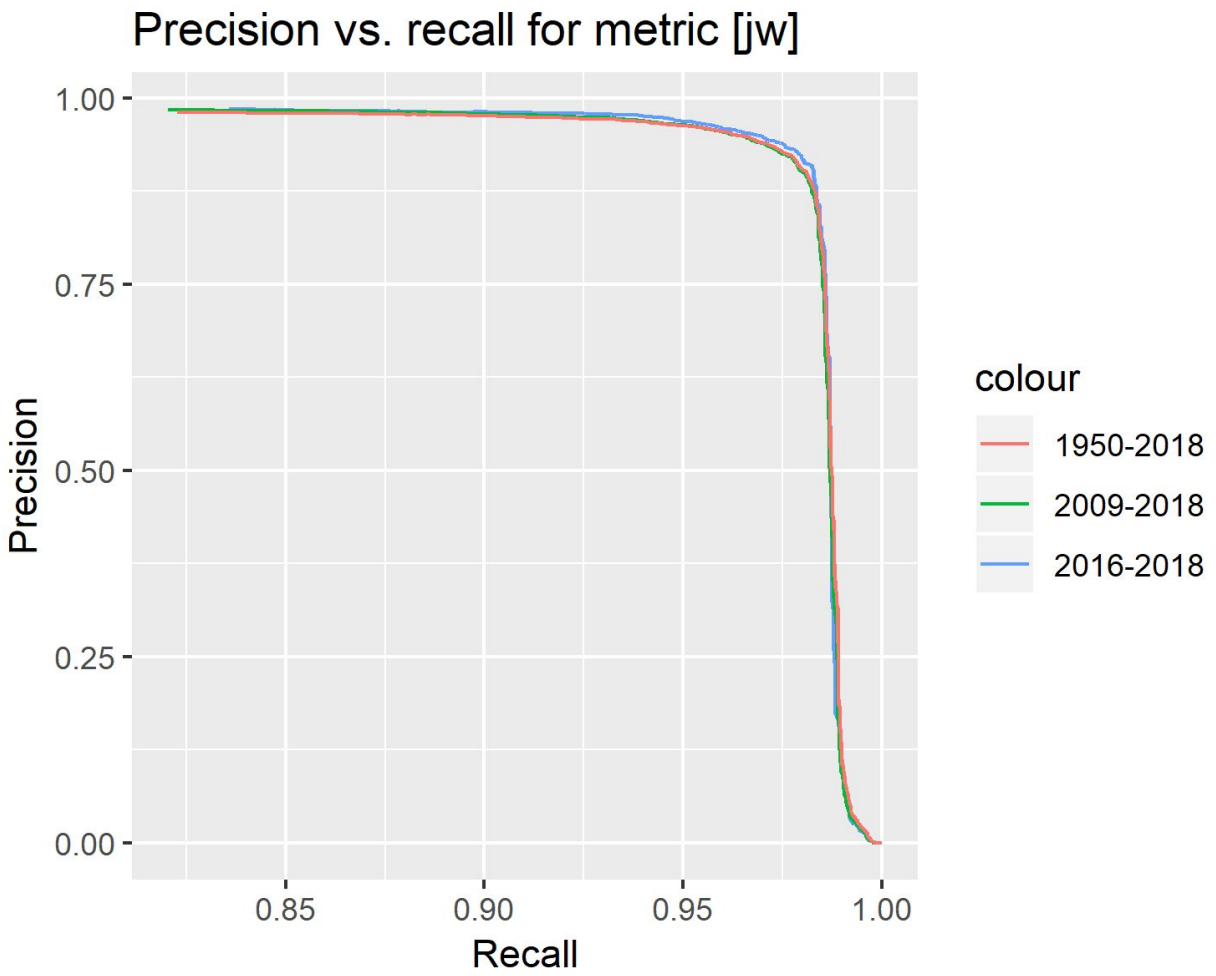
Vzhledem k podobnosti křivek bylo provedo porovnání jednotlivých časových období pro každou metriku.



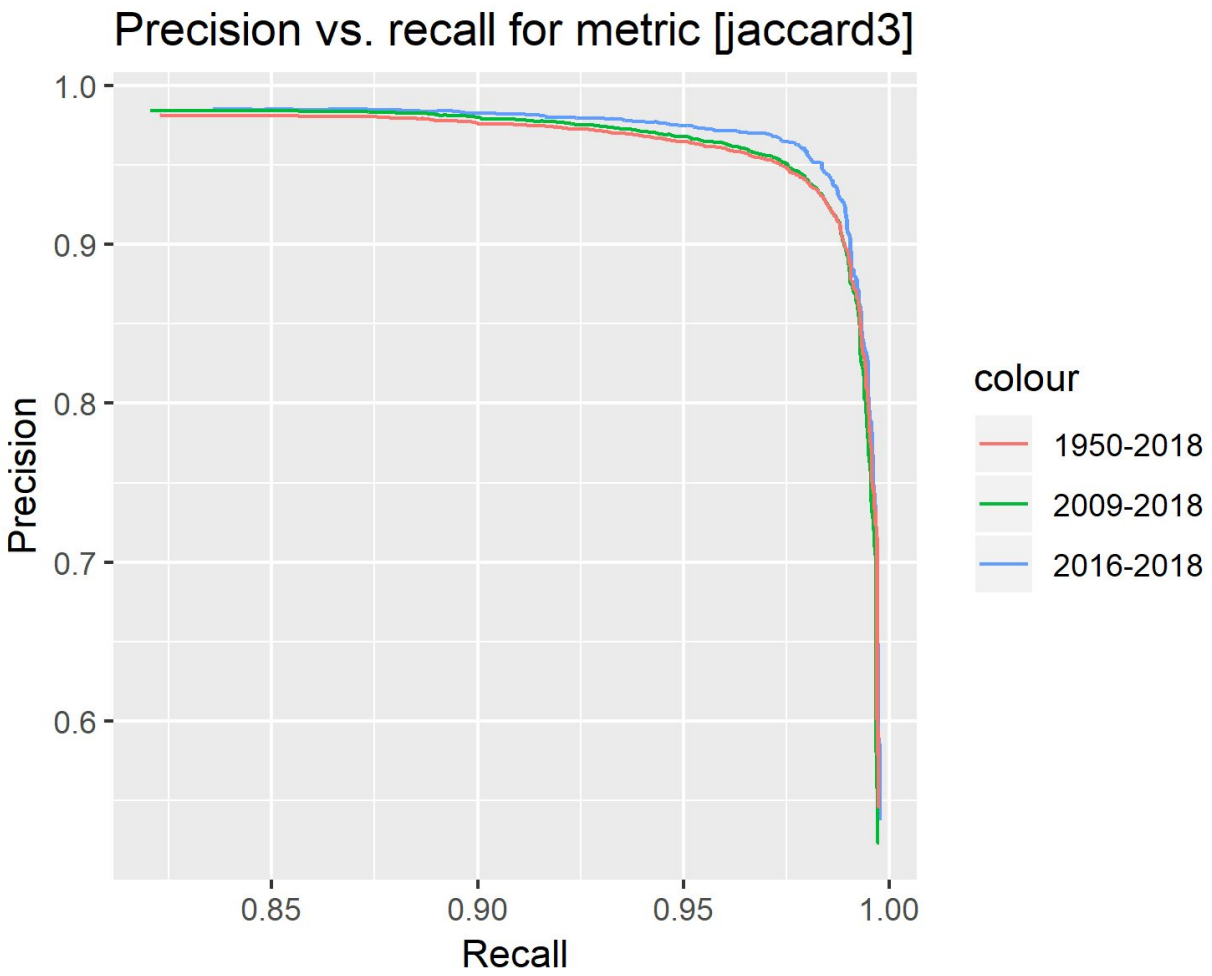
Obr. č. 20: Graf porovnávající hodnoty preciznosti a výtěžnosti Levenshteinovy vzdálenosti v jednotlivých skupinách. Publikace z let 1950-2018, 2009-2018, 2016-2018.



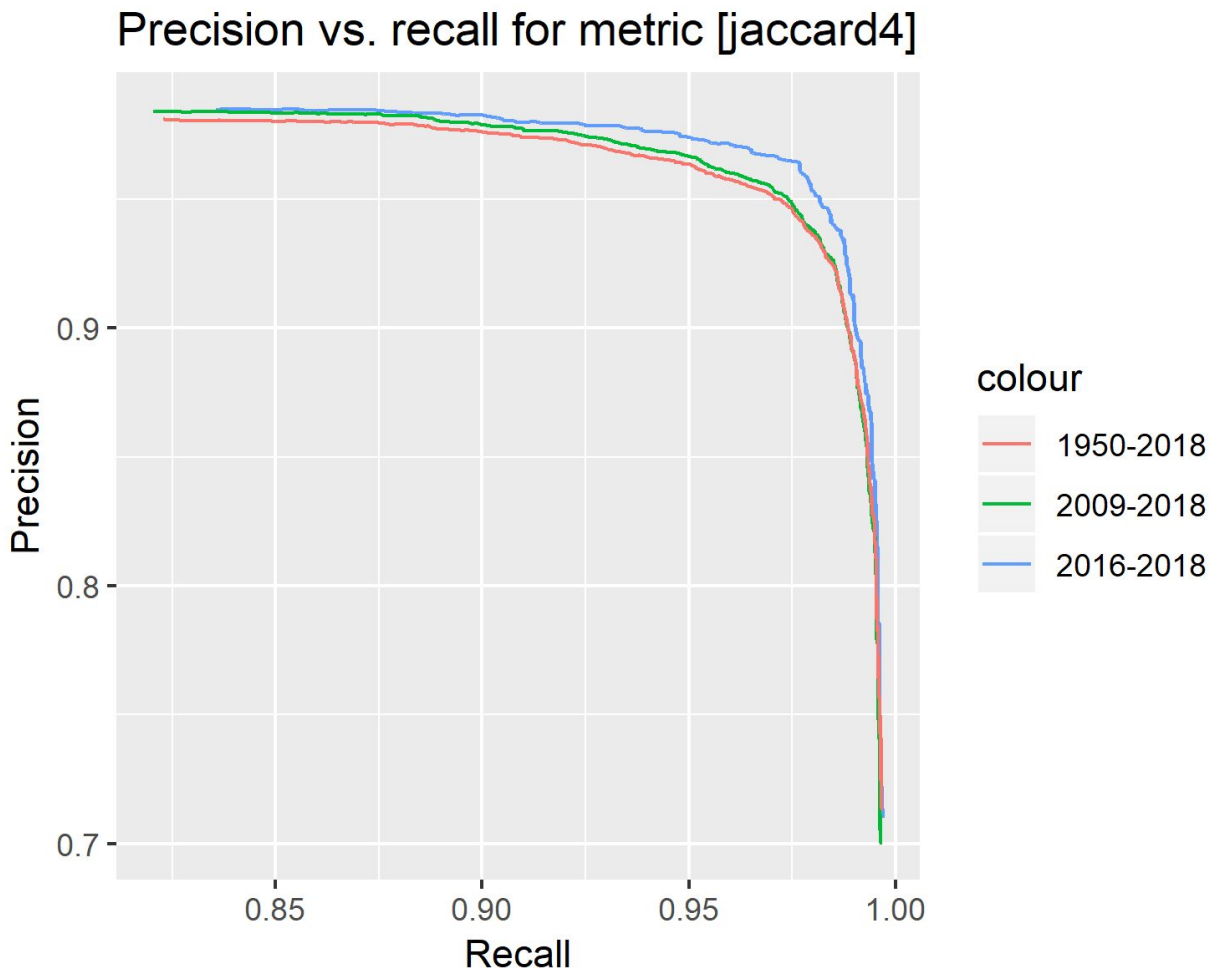
Obr. č. 21: Graf porovnávající hodnoty preciznosti a výtěžnosti Jarovy vzdálenosti v jednotlivých skupinách. Publikace z let 1950-2018, 2009-2018, 2016-2018.



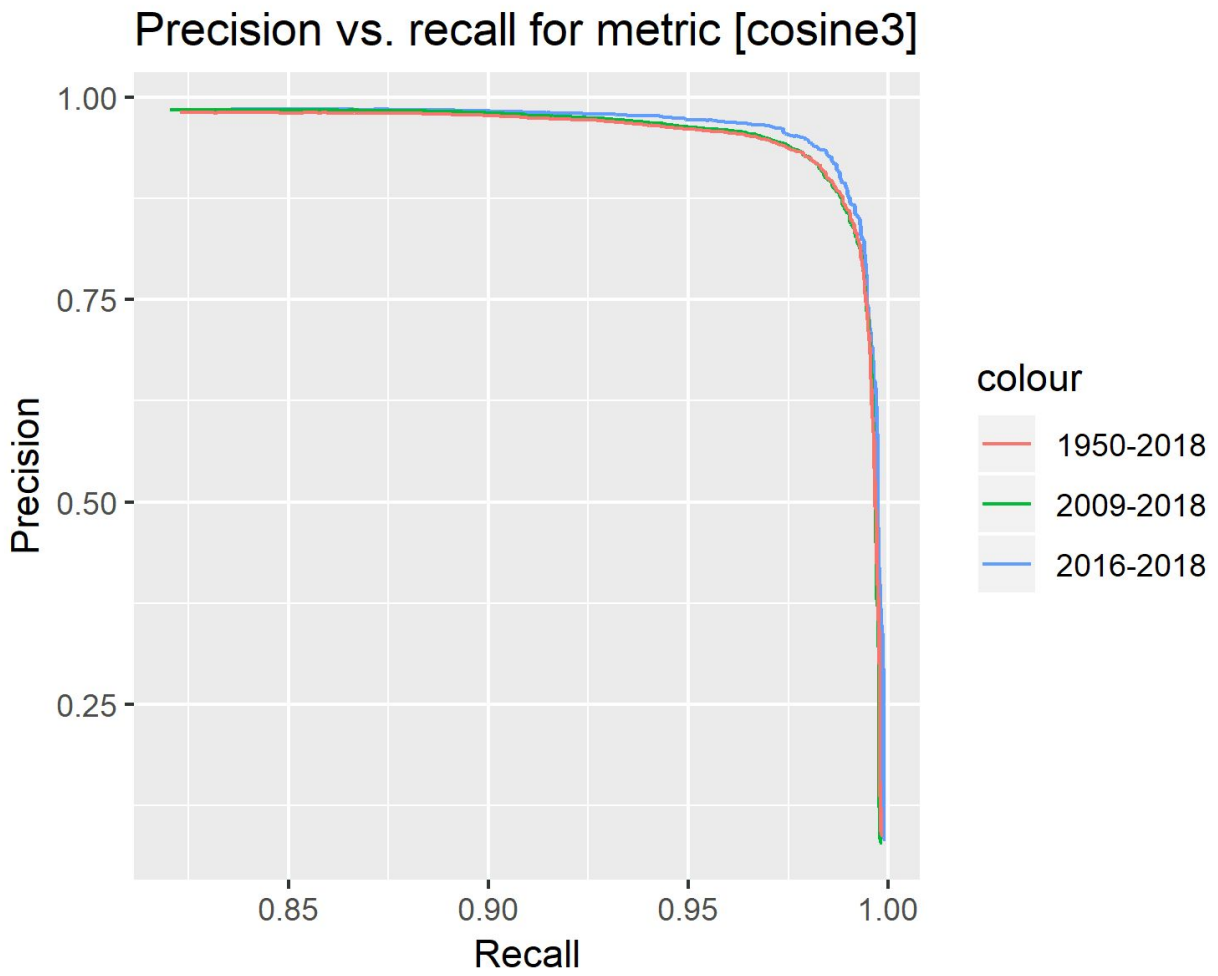
Obr. č. 22: Graf porovnávající hodnoty preciznosti a výtěžnosti Jaro-Winklerovy vzdálenosti v jednotlivých skupinách. Publikace z let 1950-2018, 2009-2018, 2016-2018.



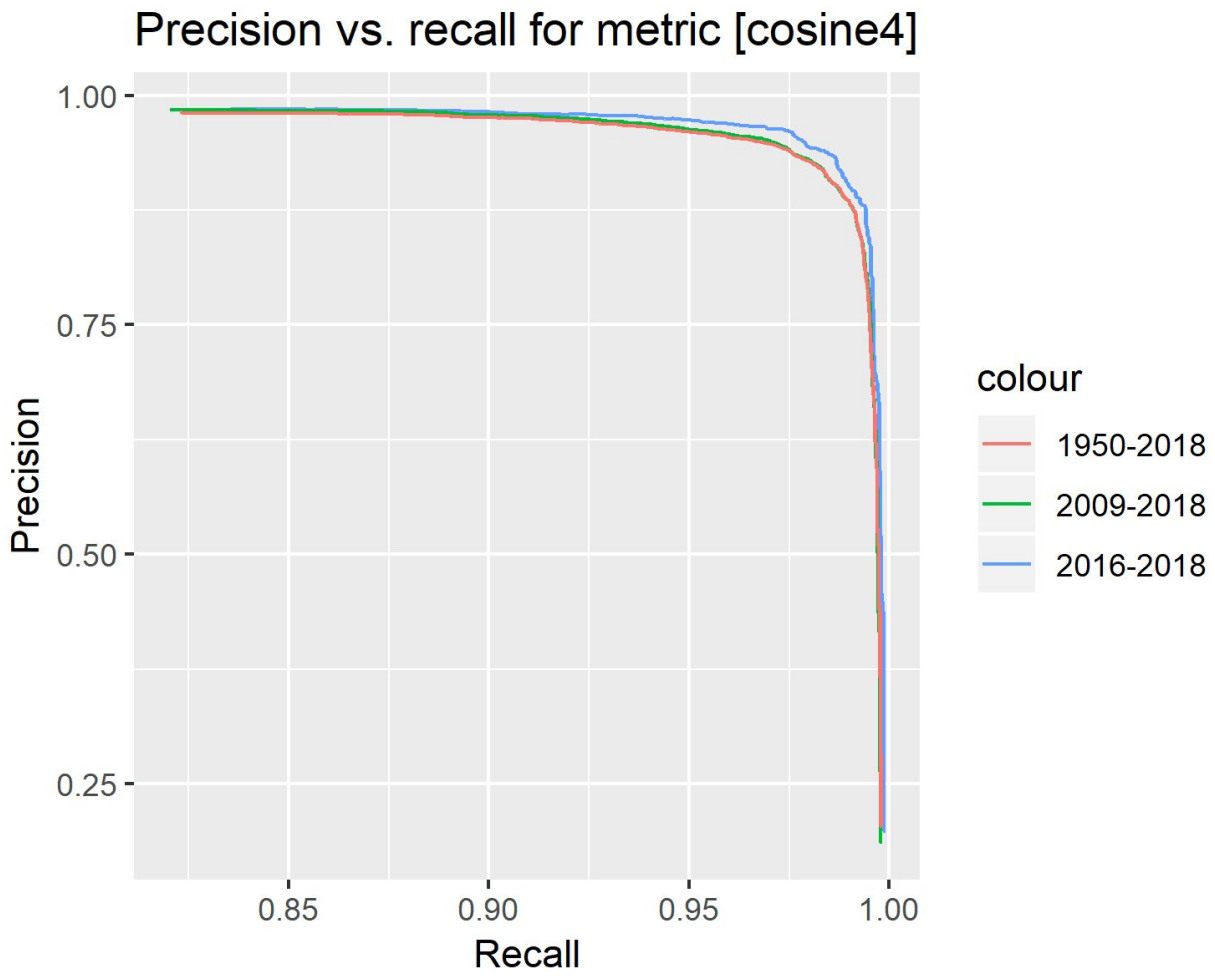
Obr. č. 24: Graf porovnávající hodnoty preciznosti a výtěžnosti Jaccardova koeficientu pro q-gramy o velikosti 3 v jednotlivých skupinách. Publikace z let 1950-2018, 2009-2018, 2016-2018.



Obr. č. 24: Graf porovnávající hodnoty preciznosti a výtěžnosti Jaccardova koeficientu pro q-gramy o velikosti 4 v jednotlivých skupinách. Publikace z let 1950-2018, 2009-2018, 2016-2018.



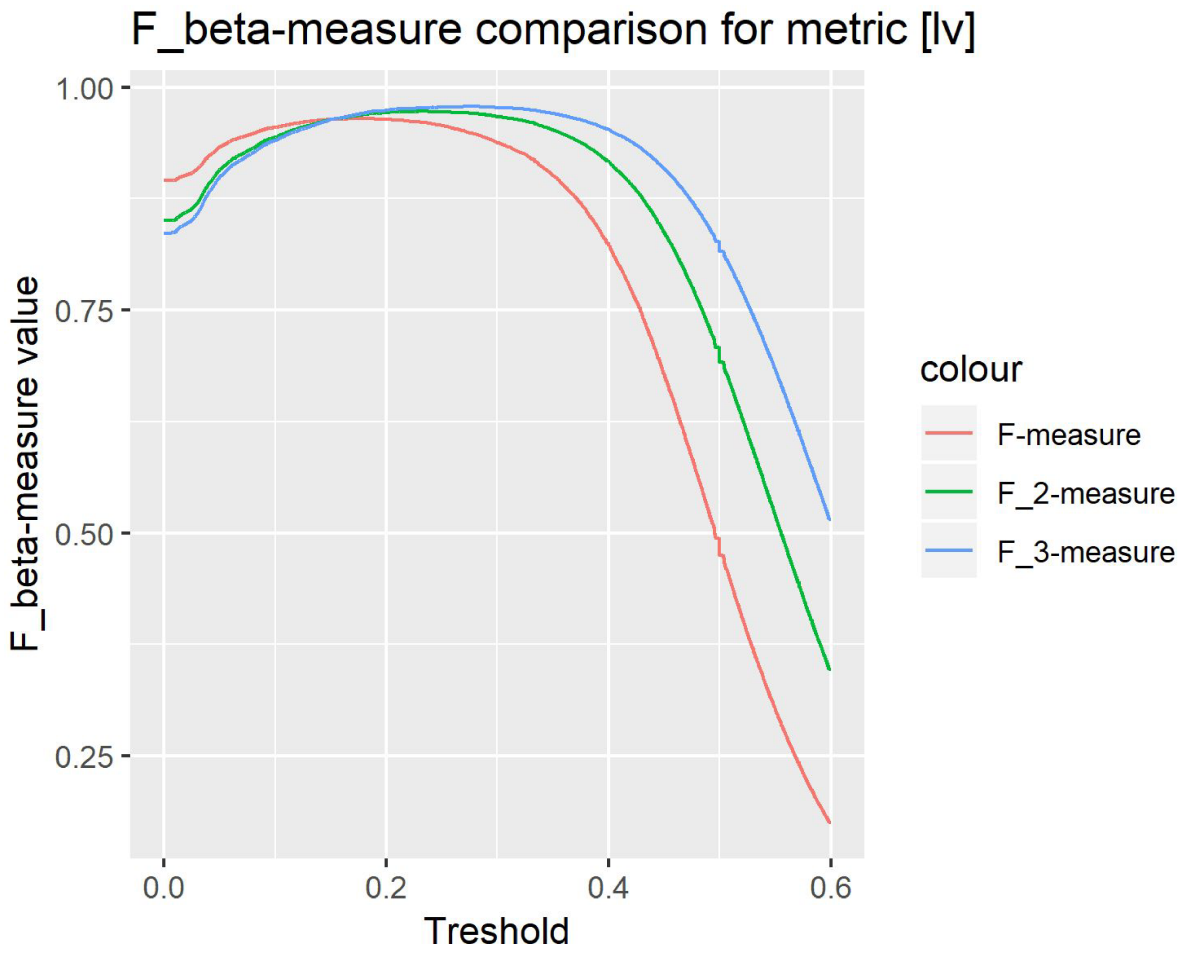
Obr. č. 25: Graf porovnávající hodnoty preciznosti a výtěžnosti kosinové vzdálenosti pro q-gramy o velikosti 3 v jednotlivých skupinách. Publikace z let 1950-2018, 2009-2018, 2016-2018.



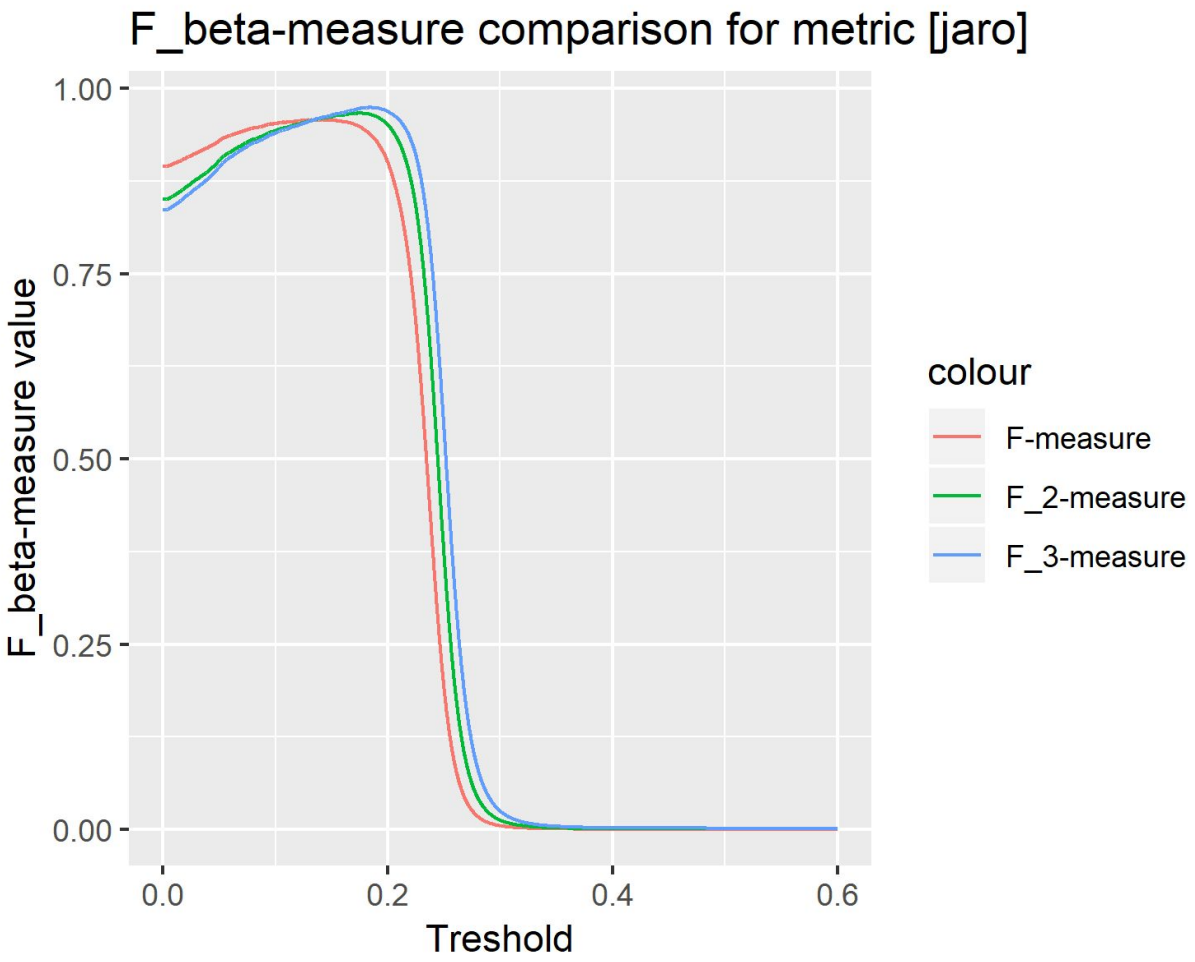
Obr. č. 26: Graf porovnávající preciznosti a výtěžnosti kosinové vzdálenosti pro q-gramy o velikosti 4 v jednotlivých skupinách. Publikace z let 1950-2018, 2009-2018, 2016-2018.

7.3 Porovnání různých F_β - měr pro jednotlivé metriky

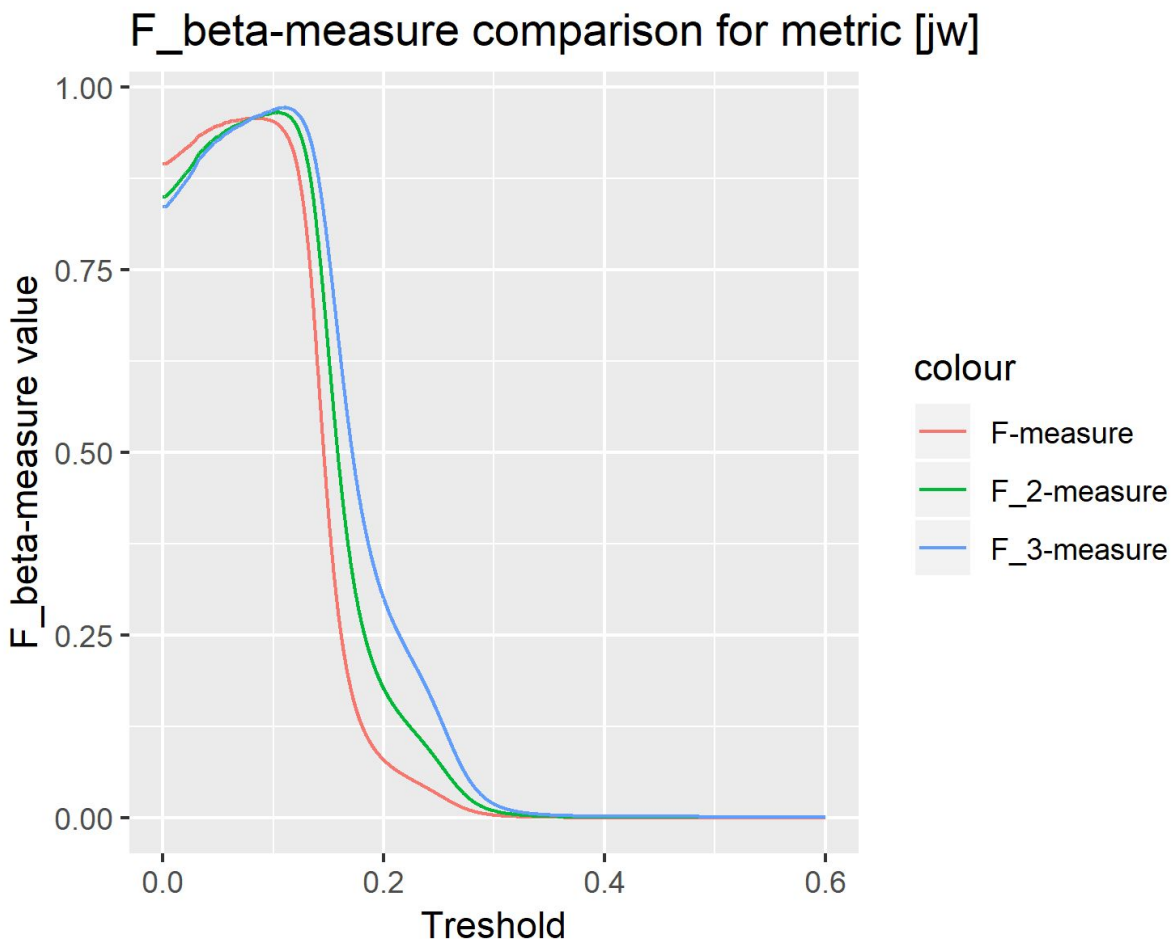
Je snazší rozpojit záznamy, které byly nesprávně propojeny, než dohledat propojené záznamy, které nebyly prahem zachyceny. Za předpokladu, že přikládáme vyšší prioritu výtěžnosti oproti preciznosti, můžeme se při hledání optimálních prahových hodnot řídit mírou F_2 a F_3 a tím zvýšit počet propojených záznamů.



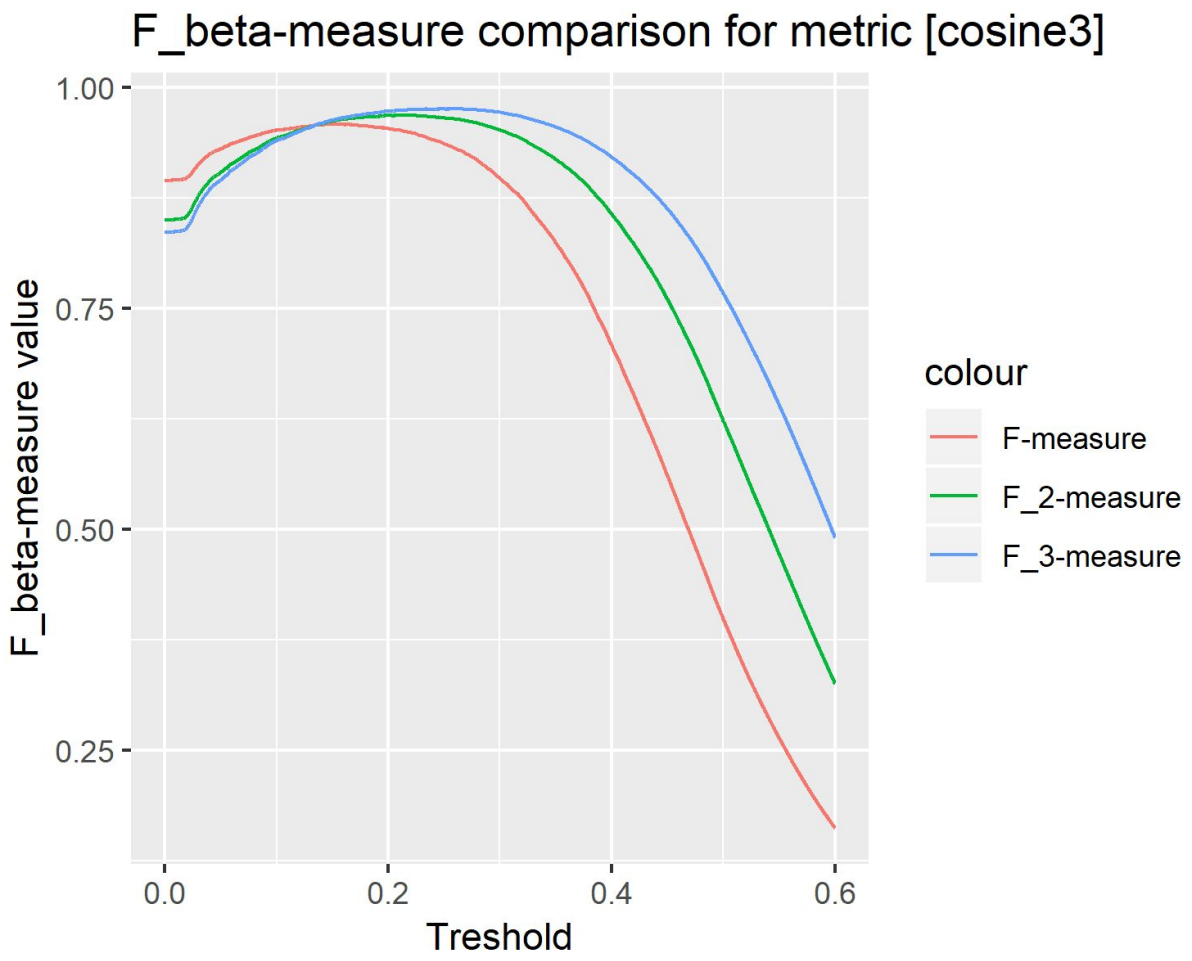
Obr. č. 27: Graf porovnávající hodnoty F_β-měr Levenshteinovy vzdálenosti na daných prazích. Publikace z let 1950-2018.



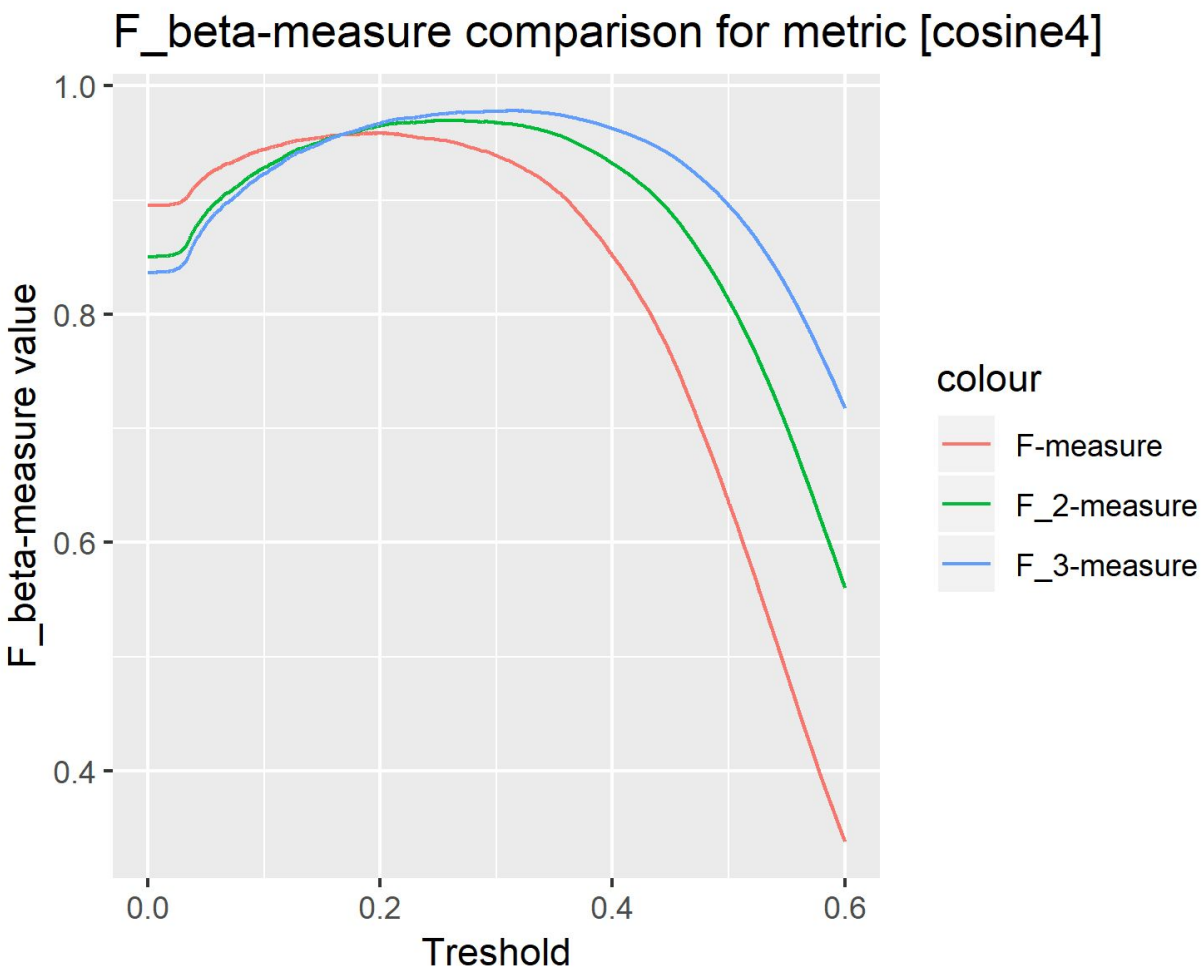
Obr. č. 28: Graf porovnávající hodnoty F_{β}-měr Jarovy vzdálenosti na daných prazích. Publikace z let 1950-2018.



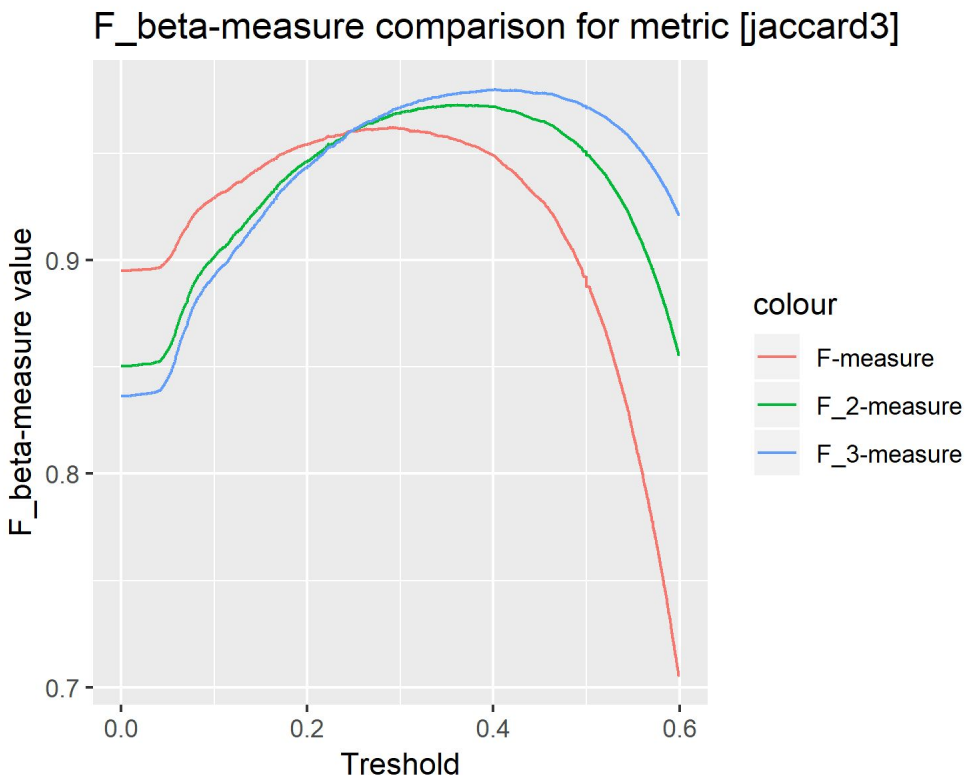
Obr. č. 29: Graf porovnávající hodnoty F_{β}-měr Jaro-Winklerovy vzdálenosti na daných prazích. Publikace z let 1950-2018.



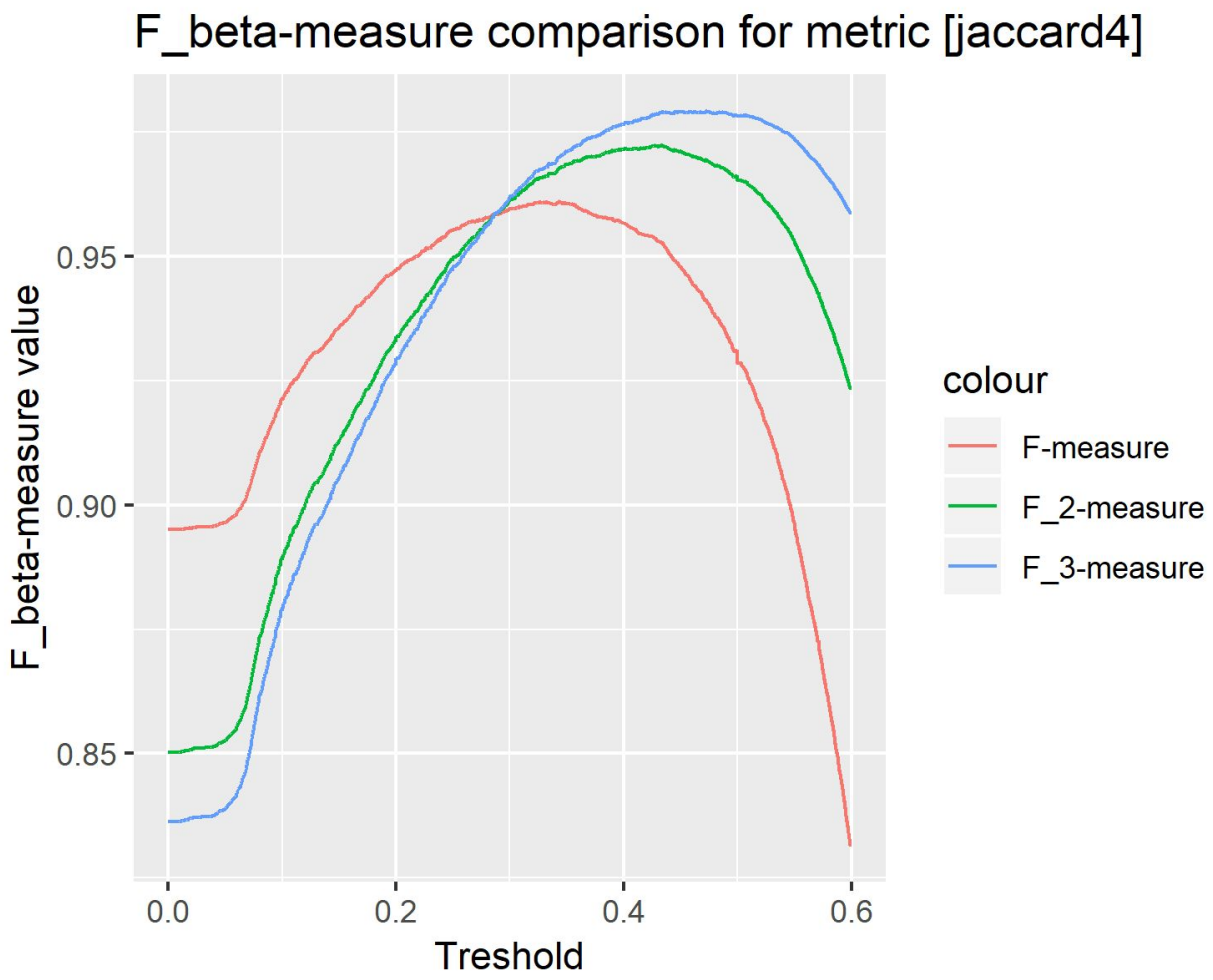
Obr. č. 30: Graf porovnávající hodnoty F_β-měr kosinové vzdálenosti pro q-gramy o velikosti 3 na daných prazích. Publikace z let 1950-2018.



Obr. č. 31: Graf porovnávající hodnoty F_β-měr kosinové vzdálenosti pro q-gramy o velikosti 4 na daných prazích. Publikace z let 1950-2018.



Obr. č. 32: Graf porovnávající hodnoty F_β-měr Jaccardova koeficientu pro q-gramy o velikosti 3 na daných prazích. Publikace z let 1950-2018.

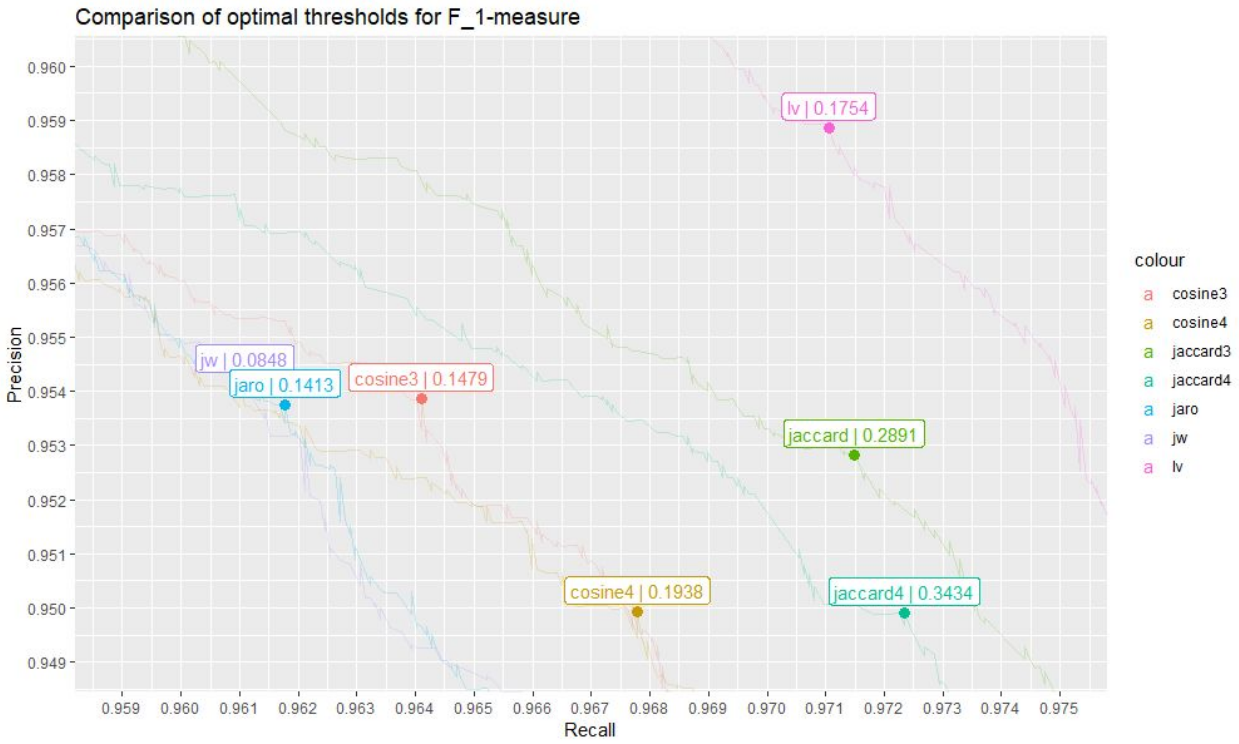


Obr. č. 33: Graf porovnávající hodnoty F_{β}*-měr Jaccardova koeficientu pro q-gramy o velikosti 3 na daných prazích. Publikace z let 1950-2018.*

Byly zjištěny následující optimální prahy (seřazeno sestupně podle hodnoty F-míry):

Optimální prahy na základě F-míry				
Metrika	**Práh**	**Preciznost**	**Výtěžnost**	**F-míra**
lv	0.1754	0.9588634	0.971053	0.964919
jaccard3	0.2891	0.9528222	0.9714826	0.9620619
jaccard4	0.3434	0.9499035	0.9723415	0.9609916
cosine3	0.1479	0.9538540	0.9640955	0.9589474
cosine4	0.1938	0.9499199	0.9677890	0.9587712
jaro	0.1413	0.9537479	0.9617763	0.9577453
jw	0.0848	0.9542043	0.9610892	0.9576344

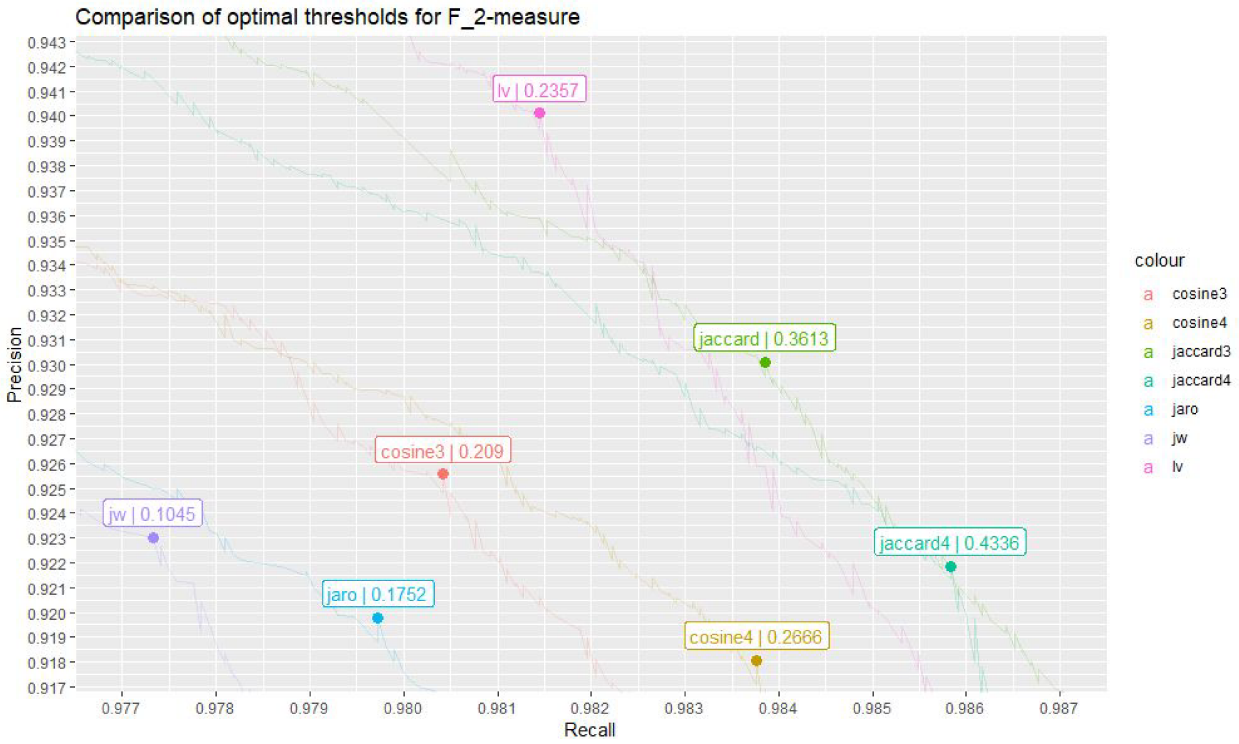
Tabulka č. 2: Optimální prahy na základě F-míry pro záznamy v letech 1950-2018



Obr. č. 34: Graf porovnávající preciznost a výtěžnost optimálních prahů. Publikace z let 1950-2018. Nejvyšší přesnosti dosahuje Levenshteinova vzdálenost, nejvyšší výtěžnost dosahuje Jaccardův koeficient pro q-gramy o velikosti 4. (Popiska jaccard má být jaccard3, pozn. autora)

Optimální prahy na základě F_2-míry				
Metrika	**Práh**	**Preciznost**	**Výtěžnost**	**F_2-míra**
lv	0.2353	0.9401020	0.9814465	0.9728892
jaccard3	0.3612	0.9300853	0.9838516	0.9726067
jaccard4	0.4335	0.9218474	0.9858272	0.9723305
cosine4	0.2666	0.9180762	0.9837657	0.9698864
cosine3	0.2090	0.9255595	0.9804157	0.9689304
jaro	0.1752	0.9197645	0.9797286	0.9671183
jw	0.1045	0.9230145	0.9773235	0.965956

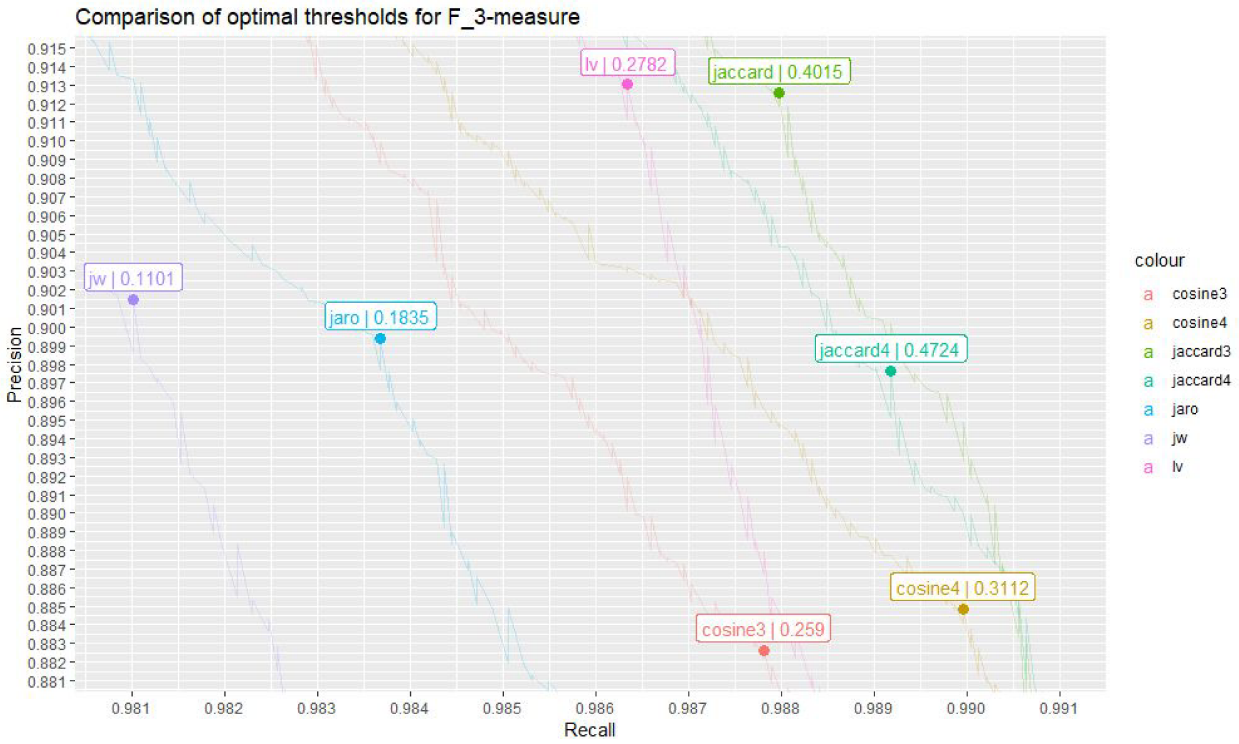
Tabulka č. 3 Optimální prahy na základě F_2-míry pro záznamy v letech 1950-2018.



Obr. č. 35: Graf porovnávající preciznost a výtěžnost optimálních prahů. Publikace z let 1950-2018. Nejvyšší přesnosti dosahuje Levenshteinova vzdálenost, nejvyšší výtěžnost dosahuje Jaccardův koeficient pro q-gramy o velikosti 4. (Popiska jaccard má být jaccard3, pozn. autora)

Optimální prahy na základě F_3-míry				
Metrika	**Práh**	**Preciznost**	**Výtěžnost**	**F_3-míra**
jaccard3	0.4015	0.9125674	0.9879746	0.9798777
jaccard4	0.4724	0.8976538	0.9891771	0.9791934
lv	0.2782	0.9130158	0.9863426	0.9784841
cosine4	0.3112	0.8848369	0.9899502	0.9783282
cosine3	0.2590	0.8826464	0.9878028	0.9761729
jaro	0.1835	0.8993953	0.9836798	0.9745471
jw	0.1101	0.9014918	0.9810170	0.9724386

Tabulka č. 4 Optimální prahy na základě F_3-míry pro záznamy v letech 1950-2018



Obr. č. 36: Graf porovnávající preciznost a výtěžnost optimálních prahů. Publikace z let 1950-2018. Nejvyšší přesnosti dosahuje Levenshteinova vzdálenost, nejvyšší výtěžnost dosahuje kosinová vzdálenost pro q-gramy o velikosti 4. (Popiska jaccard má být jaccard3, pozn. autora)

Ve všech třech případech F_β - měr dosáhla nejvyšší preciznosti Levenshteinova vzdálenost.
V případě, kdy považujeme výtěžnost za stejně, nebo dvakrát více důležitou jako preciznost, dosahuje nejvyšších hodnot výtěžnosti Jaccardův koeficient pro q-gramy o velikosti 4. Pokud bychom ale přiřadili výtěžnosti trojnásobně větší prioritu, dosahuje nejvyšší výtěžnosti kosinová vzdálenost pro q-gramy o velikosti 4.
Vzhledem k tomu, že je snažší omylem propojené záznamy rozpojit, než zpětně dohledat ty nepropojené záznamy, které by propojené být měly, lze předpokládat, že nejvhodnějšími metrikami pro ztotožňování záznamů v systému V3S by tedy byly buď kosinová vzdálenost pro q-gramy o velikosti 4, nebo Jaccardův koeficient pro stejnou velikost q-gramů. Rozhodování mezi těmito dvěma metrikami by se poté mělo odvíjet podle možností repozitáře kontrolovat množiny záznamů, jež byly chybně propojeny a jak velký by samotný počet chybně propojených záznamů byl. To by představovalo 5.1%, 7.9%, nebo až 11.6% všech porovnávaných záznamů podle nastavení priority výtěžnosti.

7.4 Výsledek hodnocení ruční kontroly

Metrika	Nepravdivých propojení v trénovacím datasetu na optimálním prahu pro F_1-measure	Celkem propojení párů ve validačním datasetu na optimálním prahu pro danou metriku	Počet nepravdivých propojení ve validačním vzorku	Typy chyb
lv	4.11%	1 022	2/100	1x odlišný chemický vzorec; 1x rozdílné konference
jw	4.58%	1 022	6/100	4x rozdíl v typu dokumentu; 1x rozdílné konference; 1x odlišný chemický vzorec
jaro	4.62%	1 023	5/100	4x rozdíl v typu dokumentu; 1x rozdílné konference
cosine3	4.61%	1 033	5/100	3x rozdíl v typu dokumentu; 1x rozdílné konference; 1x velmi podobné názvy publikací
cosine4	5.01%	1 036	3/100	1x rozdílné konference; 2x rozdílný typ publikace;
jaccard3	4.72%	1033	4/100	4x rozdílný typ publikace;
jaccard4	5.01%	1035	4/100	3x rozdílný typ publikace; 1x velmi podobné názvy publikací

Tabulka č. 5 Výsledky hodnocení ruční kontroly

Nejčastějším důvodem falešného propojení publikací byla situace, kdy autor či autoři výsledky prezentovali nejprve na konferenci a poté publikovali v odborném časopise, případně naopak. Další časté důvody jsou stejný, nebo velmi podobný název, nicméně dokument publikoval jiný autor, případně se jedná o publikaci z jiného vydavatelství. Problematický se také jeví zápis chemických vzorců, u kterých změna jednoho písmene či znaku působí velký rozdíl ve významu. Podobný efekt lze předpokládat u zkratek.

8. Diskuze

Výsledky metrik pro jednotlivá časová období mají veskrze stejný tvar, což prokazuje robustnost tohoto řešení a jeho vhodnost pro použití v systému V3S. Na některých grafech jsou vidět malé skoky v místech, kdy nebyl dostupný dostatečný počet porovnávacích párů, nicméně se nejedná o závažné zjištění. Ani v jednom případě nebylo dosaženo 100% preciznosti, což je způsobeno primárně nevalidními DOI, případně záznamy s chybně zapsaným identifikátorem. Při porovnání stejné metriky v rámci vytyčených období jsou vidět rozdíly mezi jednotlivými skupinami, zejména u menšího počtu záznamů ve skupině pro roky 2016-2018. Jednotlivé metriky nelze mezi sebou porovnávat pro jednu hodnotu prahu, mnohem cennější jsou maxima hodnot F_β - měr pro jednotlivá řešení. V systému je mnohem snazší rozpojit chybně propojené záznamy, než dohledat chybějící propojení záznamů. Proto se přikládá větší váha hodnotě výtěžnosti než hodnotě preciznosti, což vede k použití F_2-míry nebo F_3-míry. Použitím optimálních prahů pro F_2-míry a F_3-míry lze dosáhnout většího počtu pravdivě propojených záznamů.

9. Závěr

Práce představuje metody přibližné shody znakových řetězců, jež lze aplikovat v případě, kdy nejsou v metadatových záznamech dostupné jednoznačné identifikátory. Tento postup lze praktikovat i u metadat vědeckých publikací. Je popsáno několik různých metod přibližné shody znakových řetězců, jež umožňují získat hodnotu vzdálenosti znakových řetězců. Na základě vhodně vybraných metadatových polí pro ztotožňování záznamů a optimálně stanovených prahů lze posuzovat, zda-li se jedná o jednu a tutéž publikaci, nebo dvě různé publikace. Metadata vědeckých publikací se v některých případech mírně liší od běžných publikací, tyto odlišnosti však nebrání aplikaci metody.

V praktické části byly na základě dostupných dat ze systému V3S stanoveny optimální prahy pro Jaccardův koeficient, Levenshteinovu, Jarovu, Jaro-Winklerovu a kosinovou vzdálenost q-gramů. Po porovnání hodnot preciznosti a výtěžnosti na daných prazích byly doporučeny jako nejvhodnější metriky Jaccardův koeficient pro q-gramy o velikosti 4 a kosinová vzdálenost pro stejnou velikost q-gramů. Výsledky práce byly využity při implementaci funkčnosti rozpoznávání duplicit a propojování metadatových záznamů publikací pocházejících z různých zdrojů v systému V3S.

Zdrojový kód skriptu vytvořeného pro zpracování datasetu a tvorbu výsledků je dostupný na adrese https://github.com/jdobiasovsky/metric-test a dataset s výsledky byl publikován v open-access repozitáři Zenodo s DOI https://doi.org/10.5281/zenodo.3785363.

Seznam použitých zdrojů

ALVEY, Wendy a Bettye JAMERSON, 1997. Computational disclosure control for medical microdata: The Datafly system. In: *Record Linkage Techniques - 1997: Proceedings of an International Workshop and Exposition, March 20-21, 1997, Arlington, Va* [online]. B.m.: Federal Committee on Statistical Methodology, Office of Management and Budget.

BAXTER, Rohan, Peter CHRISTEN a Tim CHURCHES, 2003. A Comparison of Fast Blocking Methods for Record Linkage. In: *ACM SIGKDD '03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*. s. 2–3.

BEALL, Jeffrey, 2006. Metadata and Data Quality Problems in the Digital Library. *Journal of Digital Information* [online]. 2006, **6**(3). ISSN 1368-7506. Dostupné z: https://journals.tdl.org/jodi/index.php/jodi/article/view/65

BIERNÁTOVÁ, Olga a Jan SKŮPA, 2011. *Bibliografické odkazy a citace dokumentů: dle ČSN ISO 690 (01 0197) platné od 1.dubna 2011* [online]. Brno, 2. září 2011 [cit. 2020-04-11]. Dostupné z: https://www.citace.com/CSN-ISO-690.pdf

BRUCE, Thomas R. a Diane I. HILLMANN, 2004. The continuum of metadata quality: defining, expressing, exploiting. B.m.: ALA editions, 2004.

CAVNAR, William B. a John M. TRENKLE, nedatováno. *N-Gram-Based Text Categorization*.

COHEN, William W, 1998. Integration of heterogeneous databases without common domains using queries based on textual similarity. In: *ACM SIGMOD Record*. s. 201–212. DOI 10.1145/276304.276323

COHEN, William W, 2000. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems (TOIS)*. B.m.: ACM, **18**(3), 288–321. DOI 10.1145/352595.352598

CSE's White Paper on Promoting Integrity in Scientific Journal Publications [online]. Wheat Ridge: Council of Science Editors, 2018 [cit. 2020-04-11]. Dostupné z: https://www.councilscienceeditors.org/resource-library/editorial-policies/white-paper-on-publication-ethics/

International DOI Foundation. DOI Handbook. *doi.org* [online]. 2014 [vid. 2020-03-20]. DOI 10.1000/182

DONNER, Paul. Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics* [online]. 2017, **113**(1), 219-236 [vid. 2020-01-11]. DOI 10.1007/s11192-017-2483-y. ISSN 0138-9130.

DVOŘÁK, Jan, Tomáš CHUDLARSKÝ a Josef ŠPAČEK, 2019. Practical CRIS Interoperability. *Procedia Computer Science* [online]. B.m.: Elsevier, **146**, 256–264 [vid. 2019-12-16]. ISSN 1877-0509. DOI 10.1016/J.PROCS.2019.01.077

ERREN, Thomas C, J Valérie GROSS, Ursula WILD, Philip LEWIS a David M SHAW. Crediting animals in scientific literature. *EMBO reports* [online]. 2017, **18**(1), 18–20. ISSN 1469-221X. DOI 10.15252/embr.201643618

FELLEGI, Ivan P a Alan B SUNTER, 1969. A Theory for Record Linkage. *Journal of the American Statistical Association* [online]. B.m.: Taylor & Francis, **64**(328), 1183–1210. ISSN 0162-1459. DOI 10.1080/01621459.1969.10501049

GARDNER, Sylvia A. Spelling errors in online databases: what the technical communicator should know. *Technical Communication*, 1992, 50-53.

HARTLEY, J., 2007. There's more to the title than meets the eye: Exploring the possibilities. *Journal of Technical Writing and Communication*, **37**(1), 95–101. DOI 10.2190/BJ16-8385-7Q73-1162

HAKALA, Juha et al, 2010. Persistent identifiers: an overview. *KIM Technology Watch Report*. 2010-10-13. Dostupné z https://pdfs.semanticscholar.org/2c67/9447c394b59e095b3ef184f6e1c0f1be97fc.pdf [vid. 2020-05-03]

HAUSTEIN, S., T. D. BOWMAN a R. COSTAS. When is an article actually published? An analysis of online availability, publication, and indexation dates. In: *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference* [online]. 2015, 1170–1179. ISBN 978-975518381-7.

HERNÁNDEZ, Mauricio A. a Salvatore J. STOLFO, 1995. The merge/purge problem for large databases. In: *ACM Sigmod Record*. 127–138. DOI: 10.1145/568271.223807

HERZOG, Thomas N., Fritz J. SCHEUREN a William E. WINKLER, 2007. *Data quality and record linkage techniques*. New York: Springer Science & Business Media. ISBN 978-0-387-69502-0. DOI 10.1007/0-387-69505-2

CHRISTEN, Peter, 2012. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Transactions on Knowledge and Data Engineering* [online]. **24**(9), 1537–1555. ISSN 2326-3865. DOI 10.1109/TKDE.2011.127

CHRISTEN, Peter, 2006. A comparison of personal name matching: Techniques and practical issues. In: *Proceedings - IEEE International Conference on Data Mining, ICDM* [online]. 290–294. ISBN 0769527027. DOI 10.1109/icdmw.2006.2

CHRISTEN, Peter, 2008. Febrl--an open source data cleaning, deduplication and record linkage system with a graphical user interface (demonstration session). In: *ACM International*

Conference on Knowledge Discovery and Data Mining (SIGKDD'08), s. 1065–1068. ISBN 978-1-60558-193-4. DOI 10.1145/1401890.1402020

CHRISTEN, Peter, 2012. *Data matching : concepts and techniques for record linkage, entity resolution, and duplicate detection*. B.m.: Springer. ISBN 9783642311642. DOI 10.1007/978-3-642-31164-2

CHRISTEN, Peter a Karl GOISER, 2007. Quality and complexity measures for data linkage and deduplication. In: *Quality Measures in Data Mining. Studies in Computational Intelligence*, sv. 43, s. 127–151. Berlin, Heidelberg: Springer.
DOI 10.1007/978-3-540-44918-8_6

ISBN Users' Manual: International Edition [online]. Seventh Edition. Londýn: International ISBN Agency, 2017 [cit. 2020-04-28]. ISBN 978-92-95055-12-4. Dostupné z: https://www.isbn-international.org/content/isbn-users-manual

ISO 26324:2012. *Information and documentation — Digital object identifier system*. Švýcarsko: Maintenance Agency or Registration Authority, 2012.

JACCARD, Paul, 1912. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist* [online]. **11**(2), 37–50 [vid. 2019-12-16]. ISSN 0028-646X. DOI 10.1111/j.1469-8137.1912.tb05611.x

JARO, Matthew A., 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* [online]. **84**(406), 414–420. ISSN 1537-274X. DOI 10.1080/01621459.1989.10478785

LEVENSHTEIN, Vladimir I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady*. **10**, s. 707–710.

LOO, Mark P. J. van der, 2014. The stringdist Package for Approximate String Matching. *The R Journal* [online]. **6**(1), 111–122. DOI 10.32614/RJ-2014-011

Matušík, Zdeněk. vědecká literatura. In: *KTD: Česká terminologická databáze knihovnictví a informační vědy (TDKIV)* [online]. Praha : Národní knihovna ČR, 2003- [cit. 2020-05-04]. Dostupné z: https://aleph.nkp.cz/F/?func=direct&doc_number=000001066&local_base=KTD.

NAUMANN, Felix a Melanie HERSCHEL, 2010. An Introduction to Duplicate Detection. *Synthesis Lectures on Data Management* [online]. B.m.: Morgan & Claypool, **2**(1), 1–87. ISSN 2153-5418. DOI 10.2200/s00262ed1v01y201003dtm003

NAVARRO, Gonzalo, 2001. A Guided Tour to Approximate String Matching. *ACM Computing Surveys.* **33**(1), 31-88. DOI 10.1145/375360.375365

RAHM, Erhard a Hong Hai DO, 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* **23**(4), 3–13.

V3S - Nápověda [online] [vid. 2019-12-16]. Dostupné z: https://v3s.cvut.cz/anonymous/help

VALDERRAMA-ZURIÁN, Juan-Carlos, Remedios AGUILAR-MOYA, David MELERO-FUENTES a Rafael ALEIXANDRE-BENAVENT, 2015. A systematic analysis of duplicate records in Scopus. *Journal of Informetrics* [online], **9**(3), 570–576 [vid. 2019-12-16]. ISSN 1751-1577. DOI 10.1016/J.JOI.2015.05.002

VAN ECK, Nees Jan a Ludo WALTMAN, 2019. Accuracy of citation data in Web of Science and Scopus [online]. [vid. 2019-12-16]. Dostupné z: http://arxiv.org/abs/1906.07011

VERYKIOS, Vassilios S., George V. MOUSTAKIDES a Mohamed G. ELFEKY, 2003. A Bayesian decision model for cost optimal record matching. *The VLDB Journal*, **12**(1), 28–40. DOI 10.1007/s00778-002-0072-y

WINKLER, William E, 2006. *Overview of record linkage and current research directions.* BUREAU OF THE CENSUS [online]. Washington: U.S. Census Bureau [vid. 2019-12-16]. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.1519

WINKLER, William E, 1990. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. B.m.: The Education Resources Information Center. Dostupné z: https://www.eric.ed.gov/?id=ED325505

YAN, Su, Dongwon LEE, Min-Yen KAN a Lee C. GILES, 2007. Adaptive sorted neighborhood methods for efficient record linkage. In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: Association for Computing Machinery. s. 185–194. ISBN 978-1-59593-644-8. DOI https://doi.org/10.1145/1255175.1255213

Seznam ilustrací

Seznam tabulek